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-068-01

野生動物と旅する科学：AI 時代におけるバイオロギングの挑戦

依田 憲

Ken Yoda

名古屋大学
Nagoya University

概要

野生動物の行動や生態の研究は、精緻な野外観察を基盤として発展してきた。しかし、彼らは驚くほど容易に人間の観察可能領域を飛び越えてしまう。そのため、行動や生理的変化、さらには周辺環境との複雑な相互作用を理解するには、技術の力で「観察の限界」を突破する必要がある。こうして生まれたのが「バイオロギング（biologging）」である。アザラシやペンギンに最初の水圧ロガーが装着されてから半世紀余り。センサは飛躍的な多様化と小型化を遂げ、いまや小型種にも装着可能な軽量デバイスが実現している。バイオロギングの真価は「多視点計測」にあり、たとえば一人称視点の映像ロガーは動物の主観世界を捉え、三人称視点の GPS ロガーは広域移動を俯瞰的に描き出す。さらに、統計モデルによる多視点データの統合解析や、AI がリアルタイムに観測を制御するシステムも実用段階に入りつつあり、日本の情報科学・生態学研究者がその最前線を担っている。こうした技術融合は、ミクロ（体内の生理状態）からマクロ（渡りや回遊）へと広がる生物現象の時空間スケールを統合し、生態・神経科学・保全から工学・情報科学に至るまで、多様な分野を結びつけるハブとしてバイオロギングの可能性を拡張している。本講演では、数千キロメートルを旅する海鳥類を追跡するロマン、多視点観察がもたらす新たな生物理解、そしてウェアラブル技術やエッジコンピューティングとの親和性について、AI 研究者の皆様と共有したい。

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-068-02

音声生成 LLM を用いたサウンドスケープの進化的探索手法による生態音響指標の理解

Understanding Ecoacoustic Indices Through Evolutionary Exploration of Soundscapes Using Text-to-Audio LLMs

尾神 圭祐　有田 隆也　鈴木 麗璽

Keisuke Ogami　Arita Takaya　Reiji Suzuki

名古屋大学
Nagoya University

Abstract: 本研究は，音声生成 LLM と遺伝的アルゴリズムを組み合わせたサウンドスケープの進化的探索手法を提案し，生態音響指標の理解を深めることを目的とする．音声の時間方向の複雑さを抽出する ACI と周波数方向の複雑さを抽出する FACI，および両者の相乗平均を適応度として実験した．結果，ACI に基づく探索では生物音・非生物音を含む多様な自然音を抽出し，FACI は生物音を抽出し，両者の相乗平均は時間的・周波数的に複雑な構造を持つ生物音を抽出する傾向があることが示唆された．

1　はじめに

生態音響学（Ecoacoustics）は，ある環境における音の総体であるサウンドスケープを基に，動物・自然・人間の関係を明らかにする研究領域である[1]．自然環境のサウンドスケープは，生物が発する音（Biophony），風や川の流れといった自然由来の音（Geophony），飛行機や道路の騒音など人工的に発せられる音（Anthrophony）の３つの要素で複雑に構成されている．

自然のサウンドスケープがどのような特徴を持ち，どのような音源から構成されるかを定量的に分析するため，生態音響指標を用いた評価が行われている．生態音響指標は，サウンドスケープにおける周波数，時間，エネルギーの様々な特徴や変化を捉えるために広く利用されており，R の Seewave[2]や Python の scikit-maad[3]などのパッケージを利用して容易に計算できる．

生態音響指標には，Acoustic Complexity Index（ACI）（音響複雑性指数）という，主に鳥類の発声を定量化することを目的として考案された指標[4]や，Acoustic Diversity Index（ADI）（音響多様性指数）という，サウンドスケープの複雑さや多様性を定量的に評価する指標[5]などがある．近年，安価で長時間連続運用が可能な自動録音装置（ARU）の普及により様々な場所で録音を大量に取得可能になったため，生態音響指標に基づくサウンドスケープの特徴やダイナミクスの抽出は重要さを増しているといえる．

本研究は，その代表例である ACI に注目する．ACI は，鳥のさえずりをはじめとする多くの生物音が短期間にわたり多様な強度を示す一方で，人工的な音は比較的一定の強度を持つという経験的な観察に基づき開発された[4]．その最も単純な計算手法では，スペクトログラムの単一の周波数ビン内で隣接する２つの時間フレーム間のエネルギーの絶対差を計算し，それを全時間フレームおよび周波数帯域で合計して求められる．ACI はエネルギーの変動性や複雑さを定量化し，生物多様性の代理指標として利用されてきた．しかし，必ずしも生物多様性のみを反映しないことも示唆されている．ACI は嵐や植物から落ちる雨粒の音，昆虫の鳴き声に対しても高い値を示す場合があることも指摘されている[6]．ACI と鳥類の多様性との関係は，研究対象や環境によって一貫性を欠くことが報告されており，その生態学的解釈の難しさは課題の一つとなっている[7]．

同時に，ACI は提案者である Farina らにより複数の発展版も提案されている[8]．例えば，当初の時間軸方向の複雑性を評価する$\mathrm{ACIt_f}$に加え，周波数軸方向の複雑性を評価する$\mathrm{ACIf_t}$，および時間軸方向の音響強度の均等性を示す$\mathrm{ACIf_{te}}$等がある．これらを組み合わせることで，複雑なサウンドスケープに含まれる構成要素（野鳥の鳴声などの生物音，風や雨などの地理的物理音，航空機や交通音などの人工技術音）を自動的に検出可能にしている．例えば，$\mathrm{ACIf_t}$と

$ACIf_{te}$からなるイベント空間上の位置に応じて，遠くの合唱，豪雨，コマドリの地鳴きといった異なるイベントを識別できる．最近では，これら３指標の強度を組み合わせた音響コード分類手法も提案されており[9]，ACI は生物音から地球物理音，技術音までを包括するサウンドスケープ解析の重要なツールとしての役割が期待されている．

音声生成 AI（Text-to-Audio AI）は近年急速に発展しており，深層学習技術の登場により，より複雑で多様な音声を生成することが可能になった[10]．近年の大規模言語モデル（Large Language Model, LLM）の進展は，より柔軟で高度な音声生成を可能にし，新たな可能性を開拓している．例えば， Stable Audio[11]は，プロンプトを基に音楽制作や効果音生成など多様な用途に対応できるモデルとして期待されている．Stable Audio には，研究目的で利用が可能なオープンソース版である，Stable Audio Open が存在する．

大規模言語モデルは近年進化計算にも様々に応用されている．Meyerson らは，LLM を用いて進化計算における，Language Model Crossover（LMX）という新しい手法を提案した[12]．この手法で，LLM による交叉が，方程式，文，画像，Python コードの進化で多様な解を生成できることが示された．Suzuki and Arita は，LLM を活用した協力行動に関する性格特性の進化モデルを提案し，多様で複雑な性格特性の表現に基づいた協力行動が進化することを示した[13]．

上記で述べたように，生態音響指標はサウンドスケープの分析に広く用いられているものの，実データには様々な要因や環境の特徴が含まれるため，その値と，生物音および非生物音との対応関係の理解は十分でない．その理由の一つとして，高い ACI をもたらすサウンドスケープにどのようなものがありうるかを十分把握しきれていないことが考えられる．例えば，前述の$ACIf_t$は定義としては音響の周波数方向の複雑さを定量化するが，高い指標値をもたらす自然音がどんなものであるかは必ずしも自明でない．音声生成 AI が生み出す自然であり多様な音声を対象にしてこれを検討することは，指標の生態的意義づけの明確化に貢献しうると考えられる．

本研究は，音声生成 AI の豊かな表現力と，進化計算の持つ任意の遺伝的記述を柔軟に探索可能な能力を活用して，ありうるサウンドスケープを探索する手法を構築し，生態音響指標の理解や新しい活用への貢献の可能性を検討することを目的とする．具体的には，音声生成 LLM と LLM を用いた進化計算を利用し，特定の生態音響指標を最大化する人工的なサウンドスケープのテキストによる記述を進化的に探索する．特に，生存可能な個体の適応度に上限・下限を設け，生成時に生じる突飛な個体の独占を抑えることで探索を効率化する．分析では，探索過程で生じたサウンドスケープを，その記述と生成音声を特徴空間上の分布から分析し，対象とする指標が反映するサウンドスケープの特徴を明らかにする．

本稿では，ACI に加え，$ACIf_t$に類する周波数軸方向の変化を定量化する FACI を音響指標として定義，採用し，両者，および，両者の相乗平均を適応度とした実験を行い，提案手法の意義と可能性について検討する．

2　音声生成 LLM を用いたサウンドスケープの進化的探索手法

本研究で構築した進化的探索手法のフローチャートを図１に示す．以下，大規模言語モデルを用いた遺伝子作成や，音声生成 LLM を用いた遺伝子からサウンドスケープの生成, 生態音響指標による評価, 進化操作について説明する．

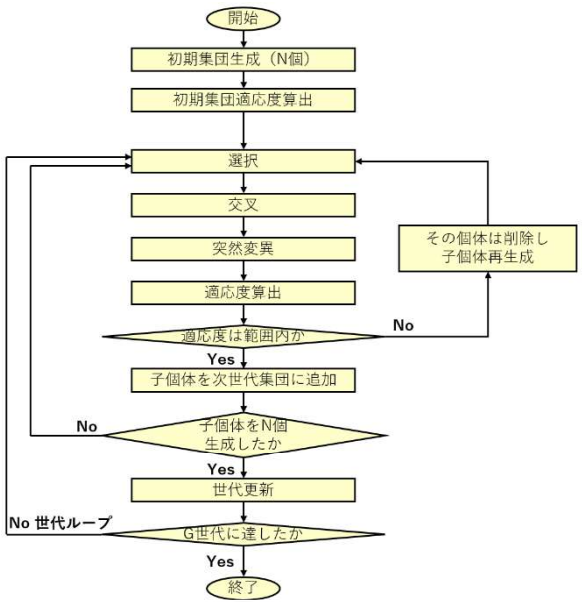


図１：進化的探索手法のフローチャート

2.1 初期集団の生成

大規模言語モデルの Llama-3.1 を使用する. 初期集団の生成は，Llama に対して以下の文章を入力し，集団サイズ N 個分の「朝の森のサウンドスケープ」を描写するプロンプトを生成した．ここで，N は生成するプロンプトの数を示す．

Create {N} one-sentence according to the theme presented below, following the rules.

Theme: "Soundscapes flowing through the forest in the morning"

Rules:

1. Create {N} descriptions.

2. There is no numbering and separate the sentences with slashes (/).

3. Each sentence must end with a period.

以下は実際に生成されたプロンプトの一部である．

The soft chirping of birds fills the air, a gentle melody.

The rustling of leaves beneath the gentle breeze is a soothing serenade.

The gentle babbling of a brook creates a peaceful ambiance.

このように，初期集団のプロンプトには，鳥のさえずりや葉のざわめき，小川のせせらぎなど，森に生息する生物や自然音について記された短い文を採用した．

2.2 プロンプトからサウンドスケープへの変換

プロンプトからサウンドスケープを表す遺伝子の表現型としての音声を生成するために，オープンソースの音声生成 AI である Stable Audio Open 1.0 を使用した[11]．このモデルは，テキストプロンプトに基づいて，最大 47 秒のステレオ音声を 44.1kHz のサンプリングレートで生成でき，テキストプロンプトから柔軟で高度な音声を生成する能力を有している．図２は，実際に「The soft chirping of birds fills the air, a gentle melody.」というプロンプトから生成された結果である．このサウンドスケープは，プロンプトの記述通り，優しく穏やかな数羽の鳥のさえずりで構成されていた．

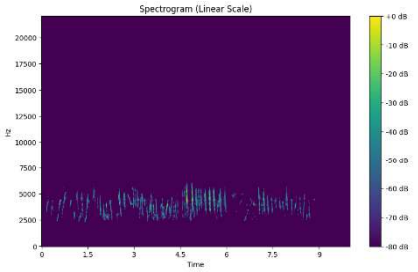


図２：実際に生成されたサウンドスケープ
プロンプト：The soft chirping of birds fills the air, a gentle melody.

2.3 生態音響指標に基づく適応度評価

音声で表される表現型の適応度として，本研究では ACI と，周波数軸に関する強度の変化量を定量化する FACI の２種類の生態音響指標を採用する．なお，今回の ACI の定義はこれが初めて提案された文献[4]での定義に基づき，FACI の定義もこれを拡張したものを採用した．ACI と FACI の計算の概要を示す図を図３に示す．

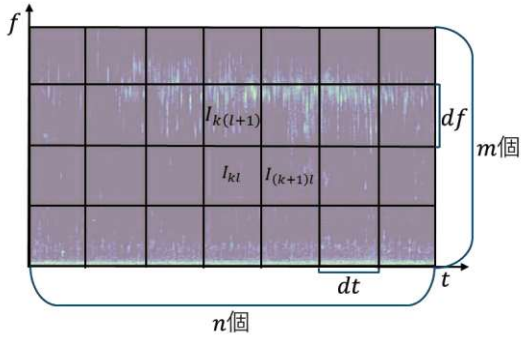


図３：ACI と FACI の計算の概要図

まず，音の時間方向の複雑性を定量的に評価できる ACI（音響複雑性指数）[4]を用いて，生物音の抽出を期待した．ACI は 次のように計算される．音声ファイルに対して短時間フーリエ変換を適用し，その時間幅dtと周波数幅dfを用いる．周波数帯域 l において，隣接する二つのタイムビンのエネルギーの絶対差d_{kl}を求める．

$$d_{kl} = |I_{kl} - I_{(k+1)l}| \tag{1}$$

ここでI_{kl}はk番目のタイムビン，l番目の周波数帯域のエネルギーを意味する．次に，その周波数帯域のd_{kl}の合計D_lを求める．

$$D_l = \sum_{k=1}^{n} d_{kl} \tag{2}$$

D_lを，その周波数帯域内のエネルギー値の合計で割って正規化した値をACI_lとする．

$$ACI_l = \frac{D_l}{\sum_{k=1}^{n} I_{kl}} \tag{3}$$

そして，全周波数帯域のACI_lを合計した値を最終的な ACI とする．

$$ACI = \sum_{l=1}^{m} ACI_l \tag{4}$$

また，Frequency-based Acoustic Complexity Index（FACI）（周波数に基づく音響複雑性指数）も適応度のための指標として採用する．鳥類のさえずりや昆虫の鳴き声などの生物音は，複雑な倍音構造と周波

数変動を持ち，特定の周波数帯域にエネルギーを集中させるのに対し．非生物音は広範な周波数帯域に連続的にエネルギーが分布する．今回採用する FACI は，このような周波数方向の複雑性と帯域特性の違いに着目することで，生物音源をより効果的に検出できる指標として機能することを期待して，前述の ACI の定義を次のように改変して定義する．まず，音声ファイルに対して短時間フーリエ変換を適用する．そのスペクトログラムについて，時間幅dtと周波数幅dfを与え，その幅でn個のタイムビン，m個の周波数帯域に再分割する．タイムビン内において，隣接する二つの周波数帯域のエネルギーの絶対差d'_{kl}を求める．

$$d'_{kl} = | I_{kl} - I_{k(l+1)} | \tag{5}$$

I_{kl}は，k番目のタイムビン，l番目の周波数帯域のエネルギーを意味する．そして，そのタイムビン内のd_{kl}の合計D_kを求める．

$$D_k = \sum_{l=1}^{m} d'_{kl} \tag{6}$$

D_kを，そのタイムビン内のエネルギー値の合計で割って正規化した値を$FACI_k$とする．

$$FACI_k = \frac{D_k}{\sum_{l=1}^{m} I_{kl}} \tag{7}$$

そして，全タイムビンの$FACI_k$を合計した値を$FACI$とする．

$$FACI = \sum_{k=1}^{n} FACI_k \tag{8}$$

FACI は, 周波数軸の振幅変動の複雑さを評価するため, 複数の生物種が異なる周波数帯域で鳴いている，すなわち生物多様性の反映を期待する．

ACI と FACI を適応度として用いるにあたり，本研究では予め設定した最小値（$ACI_{min}, FACI_{min}$）および最大値（$ACI_{max}, FACI_{max}$）を用いて Min-Max 正規化を行う．音声生成 LLM を用いた探索において指標値を無制限に最大化すると，過剰に適合した不自然なノイズが生じる恐れがあるため，自然音として妥当な範囲を上限値として設けた．また，値の範囲が異なる各指標を同一のスケールに統一することで，最適化における偏りを防ぎ，対等な評価を可能にした．正規化された値である$ACI_{norm}, FACI_{norm}$は以下の式で定義される．

$$ACI_{norm} = \frac{ACI - ACI_{min}}{ACI_{max} - ACI_{min}} \tag{9}$$

$$FACI_{norm} = \frac{FACI - FACI_{min}}{FACI_{max} - FACI_{min}} \tag{10}$$

本研究では，これらの正規化された値（ACI_{norm}，$FACI_{norm}$）の相乗平均GA_{AF}を以下の式で計算し，もう一つの適応度として採用した．これにより，時間変動のみが激しい音や単調な持続音といった片方の指標のみが高い個体の評価を低く抑え，生物音源特有の「時間的な複雑性」と，「特定の周波数帯域へのエネルギー集中」の双方を兼ね備えた個体が高い適応度を得られるよう設計した．

$$GA_{AF} = \sqrt{ACI_{norm} \times FACI_{norm}} \tag{11}$$

2.4 大規模言語モデルを用いた進化操作

集団から次世代集団を生成する際には選択と交叉，突然変異を用いる．なお，図１において世代内での操作に適応度評価が最初に存在しないのは，新規個体生成時に生存可能個体の判断の際に適応度を算出するため，計算の必要がないためである．

選択にはルーレット選択を用いた．[0, 1]の範囲に正規化された適応度を用いてルーレット選択を行い，親個体２体を選択した．これらをそれぞれ親１と親２とする．

交叉・突然変異には大規模言語モデル「Llama-3.1」を用いた．交叉は，交叉確率p_cで，Llama に対して以下の文章をプロンプト入力し，選択された親個体の遺伝子（サウンドスケープの言語記述）を組み合わせて，新たな子個体の遺伝子を生成するように指示した．プロンプト内の{parent1}および{parent2}には，親１および親２の遺伝子が入力される．もし親１と親２の遺伝子が完全に一致する場合，交叉は行わず，親１のプロンプトをそのまま子個体の遺伝子とする．交叉が適用されない場合，次の突然変異に進む．

> Create a new prompt by combining the following two original prompts according to the following rules:
> 1. The new prompt must be a single, coherent, and concise sentence.
> 2. The new prompt must be clear and to the point.
> 3. The new prompt must be 30 words or less.
>
> Original prompt 1: "{parent1}"
> Original prompt 2: "{parent2}"
> New prompt:

以下は，実際に行われた交叉の一つである．

> 親 1: A gentle hooting of an owl signals the start of a new day.
> 親 2: The soft chirping of a bird in flight creates a joyful melody.
> 子: A gentle hooting of an owl signals the start of a new day as a bird takes flight creating a joyful melody.

突然変異は，突然変異率p_mで Llama に対して以下の文章を入力し，選択された親 1 の遺伝子（プロンプト）を書き換え，新たな子個体の遺伝子（プロンプト）を生成するよう指示した．なお，プロンプト内の{original_prompt}には，親１のプロンプトが入力

される．突然変異が適用されない場合，子個体のプロンプトは親１のプロンプトと同一となる．

> Rewrite the original prompt into a new prompt according to the following rules:
> 1. The new prompt must keep the original meaning and context.
> 2. Paraphrase or incorporate a few creative variations.
> 3. The new prompt must be 30 words or less.
>
> Original prompt: "{original_prompt}"

以下は実際に行われた突然変異の一つである．

> 元のプロンプト: The soft chirping of a bird in flight creates a joyful melody.
> 突然変異後のプロンプト: Birdsong soars on gentle wings, filling the air with a sweet serenade of chirps and trills in perfect harmony.

本研究では，各指標（ACI，FACI）の値がそれぞれの最大値および最小値（ACI_{max}，ACI_{min}，$FACI_{max}$，$FACI_{min}$）の範囲外となるサウンドスケープは子個体として採用せず，新たな個体を生成しなおした．

また，計算コストを削減するため，同一のプロンプトに対応するサウンドスケープの生成をR回まで再利用する手法を採用した．具体的には，全く同一のプロンプトが与えられた場合，R回まで再利用を行い，すでにR回再利用されている場合は新たにサウンドスケープを生成する．

適応度評価と次世代集団の生成をG世代繰り返すことにより，集団を進化させた．

3　異なる適応度評価指標に基づく進化的探索の比較

本実験では，遺伝的アルゴリズム内の適応度評価において，様々な音響指標を適用し，比較分析を行う．

3.1 実験設定

表１に示すパラメータ設定を用いて, ACI, FACI, 両者の相乗平均を適応度とした条件で実験を行った．FACI の計算において, 生物音源がもつ大局的な周波数構造を捉えるため，粗い解像度である$df = 1000\text{Hz}$を採用した．一般に，風や雨などの非生物音は，広範な周波数帯域に渡ってエネルギーが連続的に分布しているが，鳥類などの生物音は種ごとに特定の周波数帯域にエネルギーを集中させる特性を持つ．したがって，ある程度広い帯域幅でエネルギーを比較し，特定の帯域に突出する生物音源の存在を効果的に検出することを期待する．

表１：実験のパラメータ

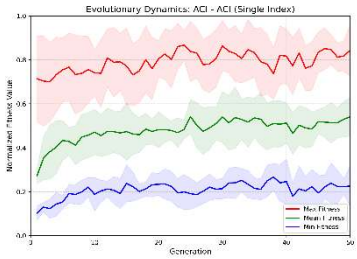
1 世代あたりの個体数	$N = 20$
世代数	$G = 50$
交叉確率	$p_c = 0.2$
突然変異率	$p_m = 0.1$
サウンドスケープの最大再利用回数	$R = 2$
各条件での試行回数	10 回
ACI の最小値	$ACI_{min} = 250$
ACI の最大値	$ACI_{max} = 500$
FACI の計算における時間の分割幅	$ds = 0.1\text{s}$
FACI の計算における周波数の分割幅	$df = 1000\text{Hz}$
FACI の最小値	$FACI_{min} = 0.0$
FACI の最大値	$FACI_{max} = 200.0$

3.2 ACI を適応度とした場合

ACI を適応度として用いた場合の最大・平均・最小適応度の推移を図 4 (a)に示す．各値は全試行の平均であり, 周囲のシェーディング領域は 10 試行間の標準偏差を示している．同図から，集団平均が世代の進行に伴い上昇しており，世代後半では安定して高い値を保っていることがわかる．これは，本手法が ACI 値に基づく適応度の高いサウンドスケープの生成を達成し，進化的探索手法として有効に機能していることを示している．

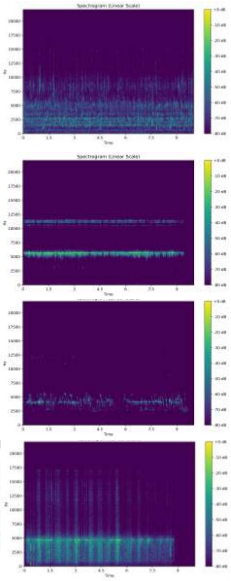
図 4 (b)は, 全 10 試行の中から, 収束した音響特性の傾向が異なる代表的な 4 試行を抽出し，その最終世代における代表個体のプロンプトとスペクトログラムを示したものである．なお，各試行の代表個体としては，適応度が中央値となる個体を選出した．これは，その世代の集団が獲得した音響的特徴を最も標準的に保持しているためである．生成された最終世代のサウンドスケープは，小川のせせらぎ（試行 2)，鳥のさえずり（試行 7, 9)，こおろぎの鳴き声（試行 7)，枝の折れる音（試行 10）など，生物音と非生物音が混在する多様な音源で構成されていた．スペクトログラムを確認すると，これらの音源はいずれも断続的で不規則なエネルギー変動を示しており，ACI の最大化という目的を達成されていることがわかる．例えば，小川のせせらぎは水流の不規則な変化として，鳥のさえずりやこおろぎの鳴き声は短時間のチャープとして，枝が折れる音は瞬間的なパルス音として現れている．

これらの結果は, ACI が鳥のさえずりだけでなく，時間的に複雑な非生物音に対しても反応するという



(a) 適応度の推移

(c) 最終世代個体の遺伝子のワードクラウド



(b) 各試行での最終世代の代表個体のスペクトログラム，遺伝子（プロンプト）と和訳．文頭の数字は試行番号.

(2) A babbling brook‘s gentle melody harmonizes with forest sounds and scents, transporting you to a serene world of nature’s symphony and tranquility.（せせらぐ小川の優しいメロディーが森の音や香りと調和し、あなたを自然の交響曲と静けさの平穏な世界へと運ぶ。）

(7) Interweave bird songs and cricket melodies to create a soothing atmosphere, evoking a sense of calm and serenity in natural settings through the harmony of nature‘s symphony.（鳥の歌とコオロギのメロディーを織り交ぜて癒やしの雰囲気を作り出し、自然の交響曲の調和を通して、自然環境における穏やかさと静けさの感覚を呼び起こす。）

(9) Birds‘ melodic trills and chirps create a harmonious symphony that fosters connection, serenity, and togetherness in nature’s beauty, bringing people together in peaceful moments of shared joy.（鳥たちの旋律的なさえずりと鳴き声が、自然の美しさの中でつながり、静けさ、そして一体感を育む調和のとれた交響曲を作り出し、共有された喜びの平和な瞬間の中に人々を引き寄せる。）

(10) Birds‘ melodic trills and chirps create a harmonious symphony that fosters connection, serenity, and togetherness in nature’s beauty, bringing people together in peaceful moments of shared joy.（活力を与える春の朝に、葉と小枝の交響曲が繰り広げられ、空気を生命と生まれ変わった季節の喜びで満たす。）

図４：ACI を適応度とした場合の実験結果

従来の知見を，進化的探索によって再現・実証したことを意味する．ACI はその計算ロジック上，時間方向の変動性を評価するため，結果として広範なサウンドスケープを生成するに至ったと考えられる．

この傾向は，図４(c)に示す最終世代個体の遺伝子（プロンプト）のワードクラウドからも見て取れる．出現頻度の高い単語には，「bird」や「birdsong」といった生物に関する語句だけでなく，「brook」や「leaves」といった自然環境や非生物に関する語句も多く含まれている．このことから，ACI の最大化は，特定の生物群に特化するのではなく，時間的に複雑なあらゆる自然音の生成を促進したことが示唆される．

3.3 FACI を適応度とした場合

次に，図４と同様にして，FACI を適応度として用いた場合の結果を図５に示す．図５(a)より，ACI の場合と同様に，集団平均が世代の進行に伴い上昇しており，世代後半では安定した高水準で収束している．この結果は，本手法が FACI 値に基づく適応度の高いサウンドスケープの生成に成功し，進化的探索手法として有効に機能していることを示している．

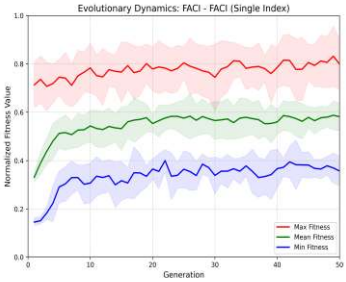
図５(b)より，生成された最終世代のサウンドスケープは，鳥のさえずりや夜の動物の鳴き声（試行 2），朝の鳥の合唱（試行 8），鳥とこおろぎの鳴き声（試行 10）といった生物音に特化しており，ACI 最大化条件で多く見られた小川のせせらぎなどの非生物音は含まれていなかった．

スペクトログラムを確認すると，ACI の解が全帯域にエネルギーが広がる傾向にあったこととは異なり，FACI の解はエネルギーが特定の周波数帯域に集中し，周波数軸方向に対して顕著な不均一性を示していることが分かる．これは，FACI の計算において 1000Hz という帯域幅でスペクトル構造を評価した設定が，広帯域に連続して分布する非生物音を抑制し，生物音源特有の特定の帯域へのエネルギー集中を効率的に抽出するフィルターとして機能したことを示唆している．つまり，FACI が周波数方向の構造に着目することで，時間方向の変動という ACI と共通の特性を持つ非生物音を効果的に排除し，生物音源を選択的に抽出したことを裏付ける結果である．

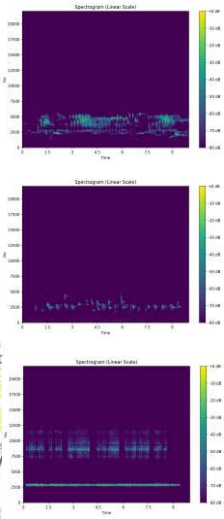
この傾向は，図５(c)に示す最終世代の個体のプロンプトのワードクラウドにも表れている．「bird」や「melodies」と言った単語が圧倒的多数を占め，「crickets」といった生物名も見られる一方で，ACI 条件で頻出した「brook」や「leaves」といった具体的な非生物音源を示す単語はほとんど見られない．このことから，FACI の最大化は，特定の生物群（特に鳥類やこおろぎ）の生成に特化するよう選択圧を与えたことが示唆される．

3.4 ACI と FACI の相乗平均を適応度とした場合

最後に，ACI と FACI の相乗平均を適応度とした場合の結果を図６に示す．図６(a)より，単一指標の条件と比較して集団平均の上昇速度が最も緩やかであることが読み取れる．これは，性質の異なる２つ



(a) 適応度の推移



(2) Under starry skies, nocturnal creatures and morning birds blend in a mystical dance with nature‘s nocturnal serenade and morning’s sweet melodies, weaving a spell of enchantment with leaves and shadows.（星空の下、夜行性の生き物と朝の鳥たちが、自然の夜のセレナーデと朝の甘いメロディーと共に神秘的なダンスで混ざり合い、葉と影で魅了する魔法を織り成す。）
(8) Awaken to a symphony of morning bird songs and golden light, where sunlight harmonizes and dances with the forest, infusing the world with serenity, joy, and vibrant peace.（朝の鳥の歌と黄金の光の交響曲で目覚めます。そこでは日光が森と調和して踊り、静けさ、喜び、そして活気に満ちた平和を世界に吹き込む。）
(10) Birds and insects blend melodies in harmony, cricketsong and forest sounds create a soothing symphony, echoing through the mist-shrouded forest, growing louder with golden light, as morning breaks.（鳥と昆虫が調和の中でメロディーを混ぜ合わせ、コオロギの歌と森の音が癒やしの交響曲を作り出し、朝が明けるにつれて黄金の光と共に大きくなりながら、霧に包まれた森に響き渡る。）

(c) 最終世代個体の遺伝子のワードクラウド

(b) 各試行での最終世代の代表個体のスペクトログラム，遺伝子（プロンプト）と和訳．文頭の数字は試行番号．

図５：FACI を適応度とした場合の実験結果

の音響特性を同時に最大化する過程が探索の難度を高めていることを示唆するものの，最終的には高い適応度へ到達しており，本手法が指標の組み合わせのような複雑な評価設定においても有効に機能していることを示している．加えて，全条件において，世代が進行しても最大適応度と最低適応度の間に一定の隔たりが維持されている．これは，集団内の解の多様性が保たれていることを示しており，進化的探索が単一の局所解に過度に集中することなく，複数の試行において安定した探索を行っていることを裏付けている．

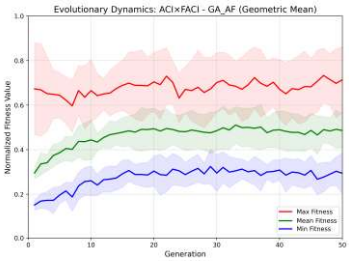
図６(b)より，生成された最終世代のサウンドスケープは，鳥のさえずりと森の情景（試行 4），鳥の合唱（試行 5），すずめの鳴き声（試行 8），こおろぎの鳴き声（試行 9）といった生物音に特化しており，ACI 最大化条件で多く見られた小川のせせらぎなどの非生物音は含まれていなかった．

スペクトログラムを確認すると，ACI と FACI の双方が指向する特徴が統合されていることが確認できる．具体的には，FACI の影響によりエネルギーが特定の周波数帯域に集中しており（広帯域ノイズの抑制），かつ ACI の影響によりその帯域内で時間的に激しく変動する構造が見られる．これは，相乗平均をとることにより，ACI が求める時間的複雑性と FACI が求める周波数構造の偏在という二つの制約が同時に課され，その結果として，両方の条件を満たす複雑な変調を持つ生物音へと解が収束したことを示している．

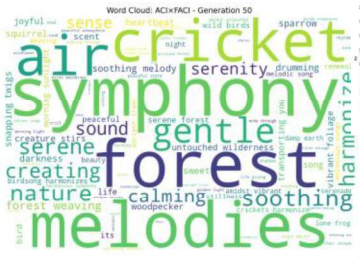
この傾向は，図６(c)に示す最終世代の個体のプロンプトのワードクラウドにも表れている．「cricket」や「melodies」が大多数を占めるほか，「sparrow」，「bird」といった生物名も見られる．一方で，ACI 単独条件で頻出した「brook」は見られず，ごく一部に「snapping twigs」が見られる程度であった．このことから，ACI と FACI の相乗平均の最大化は，非生物音を抑制しつつ，こおろぎや鳥類のような，時間的・周波数的に高度な構造を持つ生物種の生成に特化するよう選択圧を与えたことが示唆される．

3.5 進化過程の比較分析

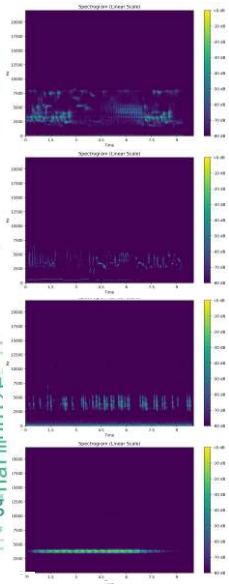
本研究で提案した進化的探索手法により，プロンプト（遺伝子）とサウンドスケープ（表現型）がどのように進化したかを視覚的に理解するため，遺伝子空間と音響特徴空間における進化軌跡を比較した．プロンプトについては，各世代の各プロンプトを Sentence Transformer によってベクトル化し，UMAP により次元削減を行って二次元空間にマッピングした．本分析では，プロンプトと主要な単語を同一空間にプロットすることで，進化が特定の意味領域へどのように引き寄せられたかを視覚的に解釈する．サウンドスケープについては，生成された音声をスペクトログラムに変換し，AST（Audio Spectrogram Transformer）の特徴量として抽出し，UMAP により二次元空間にマッピングした．なお，比較のため，次元削減は全ての条件での全試行のデータを利用し



(a) 適応度の推移



(c) 最終世代個体の遺伝子のワードクラウド



(4) Birdsong harmonizes with the fragrance of wildflowers and towering trees amidst vibrant foliage and lush greenery in a serene forest morning.（静かな森の朝、鮮やかな葉と青々とした緑の中で、鳥のさえずりが野草とそびえ立つ木々の香りと調和する。）

(5) Birds burst into enchanting melodies that captivate with stunning beauty, spreading joy and serenity as they take to the sky, their harmonious songs filling the air with vibrant energy.（鳥たちは驚くべき美しさで魅了する魅力的なメロディーをはじけるように歌い出し、空へ飛び立つ際に喜びと静けさを広げ、彼らの調和のとれた歌は空気を活気に満ちたエネルギーで満たす。）

(8) A sparrow‘s melodic song blends with the breeze, creating a serene ambiance that calms the heart with gentle trills and soothing trills.（スズメの旋律的な歌がそよ風と混ざり合い、優しいさえずりと癒やしのさえずりで心を落ち着かせる穏やかな雰囲気を作り出す。）

(9) The cricket‘s nocturnal serenade fills the stillness of night with a soothing, pulsing melody, calming the senses and transporting you to a peaceful state amidst the darkness, a constant presence.（コオロギの夜のセレナーデは、癒やしの、脈動するメロディーで夜の静寂を満たし、感覚を落ち着かせ、暗闇の中にある平和な状態へとあなたを運び、絶え間ない存在となる。）

(b) 各試行での最終世代の代表個体のスペクトログラム，遺伝子（プロンプト）と和訳．文頭の数字は試行番号．

図６ ： ACI と FACI の相乗平均を適応度とした場合の実験結果

て行い，条件ごとにデータをプロットした．

図７に，条件ごとに色分けしたプロンプト空間とサウンドスケープ空間の重心の遷移を示す．同図に示すように，全ての適応度条件において，初期集団は空間の中央付近から探索を開始している．この初期状態から，各指標に基づき世代を経るごとに集団の重心が高適応度領域へと遷移しており，提案した進化的探索手法が有効に機能していることが確認できる．

遷移の傾向を見ると，ACI を適応度とした場合 (b, f) が最も広い探索空間をカバーしているのに対し，FACI を適応度とした場合（c, g）が最も狭い空間に収束していることが示唆される．これは，ACI が評価する時間方向の複雑性が多様な領域に分散して存在している一方，FACI が評価する周波数方向の複雑性は特定の音に限定されるため，最適解の領域が狭いことを示唆している．

この差異は，適応度関数の選択が進化過程における選択圧として機能したことを意味している．ACI は幅広い自然音（生物音・非生物音の両者）に対して高い値を与えるため，解の探索空間への選択圧が弱く，拡散的になる．一方，FACI は周波数軸の構造に基づく強い選択圧により，生物音に特化した領域へと探索を集約させる．この現象は，単に高適応度解の特徴を反映するだけでなく，進化過程における選択圧が集団全体の進化方向を規定する進化ダイナミクスを示している．

ACI と FACI の相乗平均を適応度とした条件(d, h) では，両者の中間的な広さを示し，探索軌跡が安定していることから，両指標の制約が探索の方向性を絞り込み，効率化に寄与していることが考えられる．この場合，２つの異なる選択圧が均衡することで，時間方向と周波数方向の複雑性を同時に追求する「バランスの取れた」解の領域が形成されうることが示された．

4　おわりに

本研究では，音声生成 LLM と，遺伝的アルゴリズムを組み合わせた，サウンドスケープの進化的探索手法を提案した．具体的には，LLM から生成されたプロンプトを遺伝子とし，そこから生成されるサウンドスケープを表現型とする進化的探索手法を構築した．この進化的探索手法を用い，ACI，FACI，ACI と FACI の相乗平均という３つの音響指標を適応度として用い，実験を行った．その結果，全ての適応度指標において集団の適応度が明確に向上し，その生態音響指標が目標とする音響特性へ収束するプロンプトとサウンドスケープを効率的に生成できることが示された．

実験結果の分析から，生態音響指標の理解においては，単なる指標値の特徴ではなく指標が持つ選別原理を認識することの重要性が示唆された．ACI は時間的変動性を普遍的にとらえる指標であるため，自然環境の複雑さ全般を反映する可能性がある一方，特定の生物類の検出に用いる場合には限界を持つ．FACI は周波数構造に焦点を当てることで，生物音への選択性を高めるが，その結果として評価対象の空間的な広さが制限される．両者の組み合わせで，より複雑な特徴を持つ生物音の抽出の可能性が示され

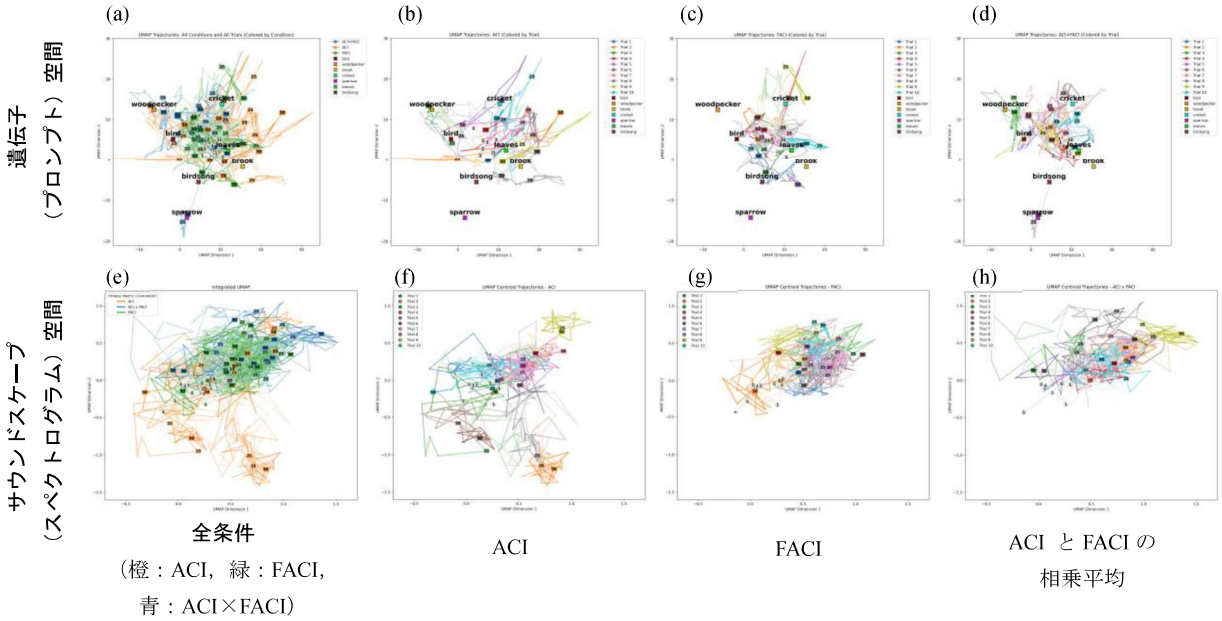


図７：全・各適応度条件における遺伝子・サウンドスケープ空間上の重心の遷移

た．

今後，より異なる指標や，複数の指標の組み合わせにより，生態学的解釈の一貫性を高めながら，サウンドスケープの多次元的な特性を捉える手法の開発が期待される．

謝辞

本研究の一部は JSPS 科 研 費 JP24K15103 の助成による．

参考文献

[1] Farina, A., Gage, S. H. (2017). *Ecoacoustics: The ecological role of sounds.* John Wiley & Sons.

[2] Sueur, J., Aubin, T., and Simonis, C. (2008). Seewave, a free modular tool for sound analysis and synthesis. *Bioacoustics*, 18(2), 213–226.

[3] Ulloa, J. S., Haupert, S., Latorre, J. F., Aubin, T., and Sueur, J. (2021). Scikit-maad: An open-source and modular toolbox for quantitative sound scape analysis in Python. *Methods in Ecology and Evolution*, 12(12), 2334–2340.

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

[5] Villanueva-Rivera, L. J., Pijanowski, B. C., Doucette, J., and Pekin, B. (2011). A primer of acoustic analysis for landscape ecologists. *Landscape ecology*, 26(9), 1233-1246.

[6] Bradfer-Lawrence, T., Gardner, N., Bunnefeld, L., Bunnefeld, N., Willis, S. G., and Dent, D. H. (2019). Guidelines for the use of acoustic indices in environmental research. *Methods in Ecology and Evolution*, 10(10), 1796–1807.

[7] Bateman, J., and Uzal, A. (2021). The relationship between the Acoustic Complexity Index and avian species richness and diversity: a review. *Bioacoustics*, 31, 614-627.

[8] Farina, A., Pieretti, N., Salutari, P., Tognari, E., and Lombardi, A. (2016). The application of the acoustic complexity indices (ACI) to ecoacoustic event detection and identification (EEDI) modeling. *Biosemiotics*, 9, 227-246.

[9] Farina, A., Righini, R., Fuller, S., Li, P., and Pavan, G. (2021). Acoustic complexity indices reveal the acoustic communities of the old-growth Mediterranean forest of Sasso Fratino Integral Natural Reserve (Central Italy). *Ecological Indicators*, 120, 106953.

[10] Moysis, L., Iliadis, L. A., Sotiroudis, S. P., Boursianis, A. D., Papadopoulou, M. S., Kokkinidis, K.-I. D., Volos, C., Sarigiannidis, P., Nikolaidis, S., and Goudos, S. K. (2023). Music deep learning: deep learning methods for music signal processing—a review of the state-of-the-art. *IEEE Access, volume 11*, 17031–17052.

[11] Evans, Z., Parker, J. D., Carr, C., Zukowski, Z., Taylor, J., and Pons, J. (2024). Stable audio open. *arXiv preprint arXiv*, 2407.14358.

[12] Meyerson, E., Nelson, M. J., Bradly, H., Gaier, A., Moradi,

A., Hoover, A. K., Lehman, J. (2023). Language Model Crossover: Variation through few-shot prompting. *arXiv preprint arXiv*, 2302.12170.

[13] Suzuki. R., Arita, T. (2024). An evolutionary model of personality traits related to cooperative behavior using a large language model. *Science Reports 14*, 5989.

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-068-03

ニューラル音場推定による 仮想音場でのマイクロフォンアレイ音源定位評価

Source Localization Evaluation of Microphone Arrays in Virtual Sound Fields Estimated by Neural Networks

加藤 雅大 [1,3]　小島 諒介 [2,3] *
Masahiro Kato [1,3]　Ryosuke Kojima [2,3]

[1] 京都大学大学院 工学研究科
[1] Graduate School of Engineering, Kyoto University
[2] 京都大学大学院 医学研究科
[2] Graduate School of Medicine, Kyoto University
[3] 理化学研究所 BDR
[3] RIKEN BDR

Abstract: Achieving sim-to-real transfer in robot audition requires accurately reproducing multichannel sound fields captured by microphone arrays, enabling tasks such as sound source localization to be performed within simulation environments. Recent sound field estimation methods, including Acoustic Volume Rendering (AVR) and Neural Acoustic Fields (NAF), have been proposed; however, most of them focus on single-channel reconstruction, and few studies evaluate sound field estimation in terms of sound source localization performance. In this work, we extend AVR and NAF to multichannel settings and evaluate them using multichannel impulse responses obtained from both real and simulated environments. The results show that the multichannel extension of NAF achieves the lowest localization error, outperforming the other methods.

1　はじめに

近年，ロボット分野においてはシミュレーション環境を活用した強化学習や模倣学習が急速に発展しており，学習した知識を実機へ転移する sim-to-real（sim2real）技術が注目されている．ロボット聴覚における sim2real を実現するためには，ロボットの物理的シミュレーションに加えて，ロボットが置かれる音環境そのものを精密に再現することが不可欠である．特に，ロボットが周囲の音源位置を把握する 音源定位（sound source localization）は，ロボット聴覚における音源追跡 [平塚 24]，行動決定 [平塚 25]，話者識別 [Mošner 24]，環境センシングにおける生態音の観測 [Heath 24] など，多様なタスクの基盤となる重要な機能である [Nakadai 20]．これらの機能は，複数のマイクから得られる信号の時空間構造を解析する マイクロフォンアレイ処理 に依存しており，現実的な音環境を再現できないシミュレーションではその性能が大きく損なわれてしまう．したがって，ロボット聴覚における sim2real の実現には，マイクロフォンアレイ処理や音源定位に対して，シミュレーション段階で現実に近い精緻な音響を再現することが重要となる．

音環境をシミュレーションするための鍵となるのが，任意の位置における音を合成できるインパルス応答（Impulse Response: IR） の利用である．ある音源信号（音声，環境音，音楽など）と IR を畳み込むことで，任意の位置で観測される音響信号を高い精度で再現できる．ニューラルネットワークを用いてインパルス応答データセットから空間全体の音場推定を行う手法として Acoustic Volume Rendering（AVR）[Lan 24] や Neural Acoustic Fields（NAF）[Luo 22] といった手法が注目されている．しかし，これらの手法は単一チャンネルの音響データを対象にしていることが多く，多チャンネルデータに注目した手法は限られている．

そこで，本研究では，実環境及びシミュレーション環境において多チャンネルのインパルス応答データセットを構築し，AVR や NAF を多チャンネルデータに拡張した手法を用いて音場推定を行う．また，各手法で

*連絡先：京都大学
京都府京都市左京区聖護院川原町 54
E-mail:kojima.ryosuke.8e@kyoto-u.ac.jp

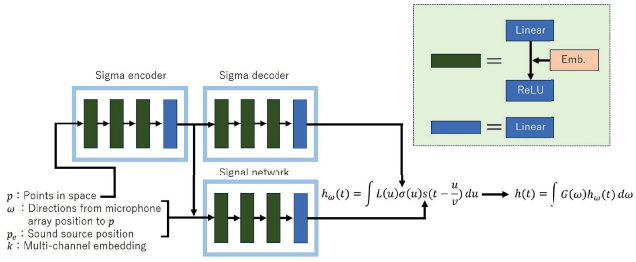


図 1: 多チャンネル AVR アーキテクチャ

生成したインパルス応答を用いて音源定位を行い，その誤差により各音場推定手法の評価を行う．本実験のコードは GitHub[1] より利用可能である．

2 ニューラル音場推定手法

本研究では，ニューラルネットワークを用いた音場推定手法として，音源位置とマイク位置の座標を入力とし，それに対応したインパルス応答を推定する手法である AVR 及び NAF をベースとし，それらを多チャンネル拡張した手法を提案する．

2.1 AVR ベースの手法

Acoustic Volume Rendering（AVR）[Lan 24] は，画像処理における複数の 2 次元画像から 3 次元シーンを再構成する手法の NeRF [Mildenhall 20] における Volume Rendering の考え方を音響信号に拡張し，位相整合性を保ちながらインパルス応答を生成する手法である．AVR では，マイク位置を中心とした球面上に放射される方向 ω を考え，その方向上の空間点で観測される局所的な音響信号をニューラルネットワーク $F : (p, \omega, p_e) \mapsto (\sigma, s(t))$ により表現する．

ここで，p は方向 ω に沿った空間中の点，p_e は音源位置，σ はその点における音響密度，$s(t)$ は t を時刻として，点 p からマイク方向へ伝搬する局所的な音響信号である．

次に，マイク位置から方向 ω に u だけ離れた点 $p(u) = p + u \cdot \omega$ について，密度 σ や伝搬距離に応じた遅延，減衰を考慮しながら $s(t)$ を u_n から u_f の範囲で積分することで，方向 ω からマイクに到達する信号 $h_\omega(t)$ は以下のように表現できる．

$$h_\omega(t) = \frac{1}{t_v}\int_{u_n}^{u_f} L(u)\,\sigma(p(u))\; s\Big(t - \frac{u}{v};\, p(u), \omega\Big)\; du,$$

$$\text{where} \quad L(u) = \exp\Big(-\int_{u_n}^{u} \sigma(p(x))\,dx\Big).$$

ただし，v は音速，t_v は距離によるエネルギー減衰，$L(u)$ は透過率を表す．

最後に，マイクの指向性 $G(\omega)$ を考慮して $h_\omega(t)$ を各方向 ω について積分することで，マイク位置でのインパルス応答 $h(t)$ を得る．

$$h(t) = \int G(\omega)\, h_\omega(t)\, d\omega.$$

本稿では，AVR の多チャンネル拡張を AVR+と呼び，以下のように拡張する．図 1 に AVR+の全体のアーキテクチャを示した．AVR+では，上述のニューラルネット F の代わりに，チャンネル番号 k を入力に含む次のニューラルネット $F' : (p, \omega, p_e, k) \mapsto (\sigma_k, s_k(t))$ を使用する．ここで，チャンネル番号 k は埋め込みベクトルとして，ニューラルネットに入力する．

さらに，追加の拡張をした手法を AVR++と呼び，音源定位に関する誤差項を損失関数に加える．具体的には，微分可能性と計算量を考慮して比較的単純な delay-and-sum beamforming を用いた損失関数を定義する．ここで，マイクロフォンアレイ中心から k 番目のマイク位置へのベクトルを $\mathbf{p}_k$ として，方向 $\mathbf{u}(\theta_\ell)$ に対するパワースペクトル P_ℓ は，次のように表される．

$$P_\ell = \sum_f \left(|A_{\ell,f}|^2 \Big/ \sum_{\ell'=0}^{L-1} |A_{\ell',f}|^2 \right),$$

$$A_{\ell,f} = \frac{1}{K}\sum_{k=1}^{K} X_{k,f}\, \exp\Big(-j2\pi f\, \frac{\mathbf{p}_k^\top \mathbf{u}(\theta_\ell)}{v}\Big),$$

$$\mathbf{u}(\theta_\ell) = \begin{bmatrix} \cos\theta_\ell \\ \sin\theta_\ell \end{bmatrix}, \qquad \theta_\ell = \frac{2\pi\ell}{L}.$$

$X_{k,f}$ は k 番目のマイクで観測された信号をフーリエ変換により周波数領域へ変換したときの，周波数ビン f の複素スペクトルである．計算したパワースペクトル P_ℓ を用いて，微分可能となるように soft-argmax を用いて音源定位方向 $\hat{\theta}$ を計算する．

$$\hat{\theta} = \sum_{\ell=0}^{L-1} w_\ell\, \theta_\ell, \qquad w_\ell = \frac{\exp(\beta P_\ell)}{\sum_{\ell'=0}^{L-1} \exp(\beta P_{\ell'})}.$$

正解波形及び AVR 推定波形による音源定位方向をそれぞれ $\hat{\theta}_{\text{true}}$，$\hat{\theta}_{\text{pred}}$ として，次のように計算した誤差 L_{DS} を従来の AVR の損失関数に加えて，ニューラルネットワークの学習を行う．

$$\mathcal{L}_{\text{DS}} = \left|\sin\hat{\theta}_{\text{pred}} - \sin\hat{\theta}_{\text{true}}\right| + \left|\cos\hat{\theta}_{\text{pred}} - \cos\hat{\theta}_{\text{true}}\right|.$$

ここでは，音源候補方向の分解能 L は 360 とした．

[1] https://github.com/KMASAHIRO/multichannel-soundfields

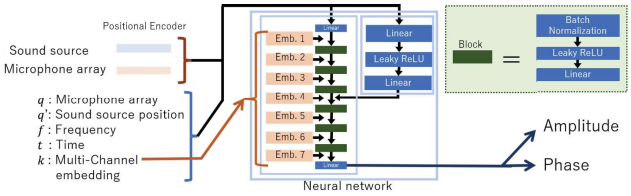


図 2: 多チャンネル NAF アーキテクチャ

2.2 NAF ベースの手法

Neural Acoustic Fields（NAF）[Luo 22] は，時間-周波数領域のインパルス応答を予測対象として，マイク位置 q，音源位置 q'，時刻 t，周波数 f に対応するスペクトルの振幅 v_{mag} 及び位相 v_{IF} を，ニューラルネットワーク $\Omega : (q, q', t, f) \mapsto \big(v_{\mathrm{mag}}(t, f), v_{\mathrm{IF}}(t, f)\big)$ により推定する．

本稿では，NAF の多チャンネルデータ拡張を NAF+ と呼び，ニューラルネットワーク Ω の代わりに，チャンネル番号 k の埋め込みベクトルを利用するニューラルネットワーク $\Omega' : (q, q', t, f, k) \mapsto \big(v_{\mathrm{mag},k}(t, f), v_{\mathrm{IF},k}(t, f)\big)$ を使用する．

NAF+では従来の 2 チャンネルに対する NAF のアーキテクチャをベースに，図 2 に示すアーキテクチャを構成した．ここで，各チャンネルの埋め込みベクトルは中間層の次元と同じ次元数を持つ埋め込みを設定して，各層の間に足し合わせる構造となっている．

3 データセット

ここでは，実環境での測定により構築したデータセット，AcoustiX[Lan 24] 及び Pyroomacoustics[Scheibler 18] によるシミュレーションで構築したデータセットの計 3 つの評価データセットについて述べる．

3.1 実データセット

実環境における TSP (Time Stretched Pulse) 信号を用いた計測によって，実データの多チャンネルインパルス応答データセットを構築した [加藤 24]．データセットは，同一環境でのマイク位置，音源位置，多チャンネルインパルス応答をデータの一組として，合計 184 個の組から構成される．

次に，具体的な測定条件を示す．図 3 および図 4 に示す環境で，部屋の寸法は 6.110m（幅方向）×8.807m（奥行方向）であった．スピーカーと円形マイクロフォンアレイを 1 つずつ用意し，幅方向に 1m 間隔で 4 箇所，奥行方向に 1m 間隔で 6 箇所の計 $4 \times 6 = 24$ 箇所をマイクロフォンアレイの配置候補箇所とし，角の 4 箇所及び中央の 4 箇所の計 8 箇所をスピーカーの配置候補箇所とした．スピーカー及びマイクロフォンアレイの高さは 1.5m に固定した．円形マイクロフォンアレイにはチャンネル数は 8，半径は 0.0365m の TAMAGO (SYSTEM IN FRONTIER Inc.) を使用した．

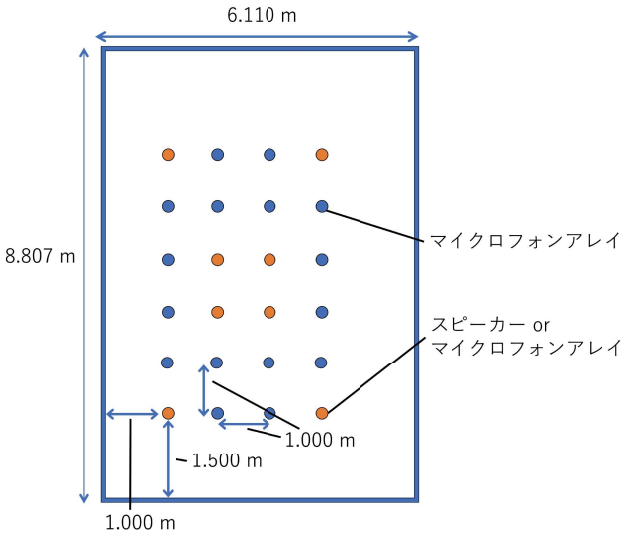


図 3: データ測定環境

図 4: 実データ測定時の部屋の写真

計測時のサンプリング周波数は 16kHz とし，TSP 信号の再生及び記録には Audacity[2]を用いた．スピーカー及び円形マイクロフォンアレイの位置を変えて測定を繰り返し，計 $8 \times 23 = 184$ 回の測定を行った．1 回の測定に付き，スピーカーから TSP 信号を連続で 10 回再生させ，円形マイクロフォンアレイで記録を行った．

測定終了後，記録した 10 個の TSP 信号をそれぞれ ITSP 信号と畳み込んでインパルス応答を生成し，10 個のインパルス応答の平均を取った．その後，測定間の記録開始タイミングのずれを調整するため，1 番目チャンネルのインパルス応答における振幅のピーク時刻が全データで一致するように加工した．また，振幅

[2] https://www.audacityteam.org/

のピーク値の全データ間での最大値により各インパルス応答を正規化した．さらに，AVR 提案論文の状況に合わせて，各インパルス応答の長さを 0.1s に切り詰めた．このようにして，計 184 個の多チャンネル室内インパルス応答からなる実データのデータセットを得た．

実データセットのクオリティチェックとして，幾何計算から求めた伝達関数による周波数正規化 MUSIC 法 [Salvati 14] によって，収録信号を用いた音源定位を行ったところ，実際の音源方向から 6.50° ± 4.99° の絶対誤差があった．

3.2 シミュレーション 1 (AcoustiX)

AcoustiX[Lan 24] は，レイトレーシングを用いた音響シミュレータである．電波用のレイトレーシングエンジンである Sionna ray tracing（Sionna RT）[Hoydis 22] を音響向けに拡張したシミュレータであり，Sionna RT が計算した各レイの反射経路に対して遅延時間や減衰を考慮することで，位相や到達時刻の正確性に注意したインパルス応答の計算が可能である．

AcoustiX を用いて，3.1 節と同様の設定（部屋寸法，音源及びマイクロフォンアレイ配置，マイクロフォンアレイ形状）でシミュレーションを行った．シミュレーションの設定値はデフォルトの値を用い，レイの数は 50000，反響回数 10 回まで追跡し，サンプリング周波数は 16kHz，0.1s のインパルス応答を計算した．このようにして，計 184 個の多チャンネル室内インパルス応答からなるシミュレーションデータセット 1 を得た．

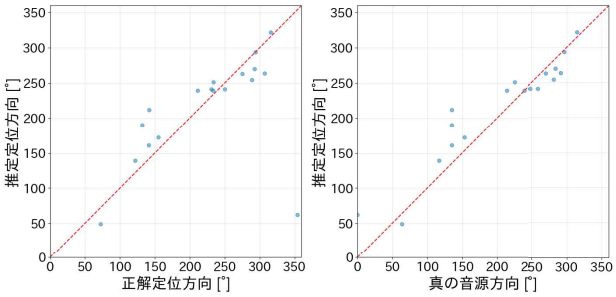
クオリティチェックとして，周波数正規化 MUSIC 法を用いてシミュレーション 1 の信号から音源定位を行ったところ 3.11° ± 3.67° の絶対誤差があった．

3.3 シミュレーション 2 (Pyroomacoustics)

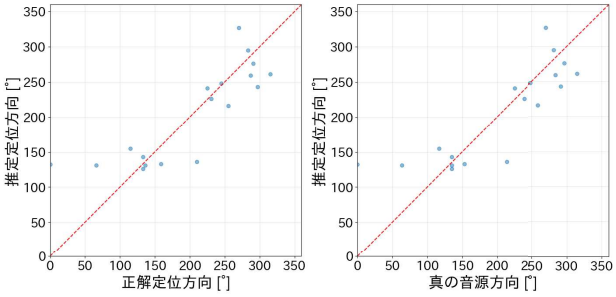
Pyroomacoustics[Scheibler 18] は広く利用される音響シミュレータであり，鏡像法を用いて効率的にインパルス応答の計算が可能である．

Pyroomacoustics を用いて，3.1 節及び 3.2 節と同様の設定（同一部屋寸法，音源及びマイクロフォンアレイ配置，マイクロフォンアレイ形状）でシミュレーションを行った．反響回数は 10 回まで追跡し，サンプリング周波数は 16kHz，0.1s のインパルス応答を計算した．このようにして，計 184 個の多チャンネル室内インパルス応答からなるシミュレーションデータセット 2 を得た．

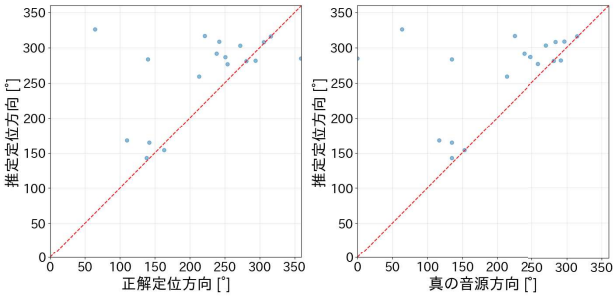
クオリティチェックとして，周波数正規化 MUSIC 法を用いてこのシミュレーションデータセット 2 の信号から音源定位を行ったところ 5.83° ± 6.46° の絶対誤差があった．



(a) 実データ



(b) シミュレーション 1



(c) シミュレーション 2

図 5: 実験 1 の結果，左図：DoA-A 誤差に対応（横軸：正解定位方向，縦軸：推定定位方向），右側：DoA-B 誤差に対応（横軸：真の音源方向，縦軸：推定定位方向）

4 実験

4.1 実験 1: ニューラル音場推定手法の比較

2 節で提案した各手法の比較をするため，3 節で構築した各データセットに対し，90%を訓練データ，10%を検証データとして分割し，推定インパルス応答を入力として周波数正規化 MUSIC 法を用いた音源定位誤差による各手法の評価を行った．各手法は Optuna を用いて，それぞれ表 3 と表 2 の範囲でハイパーパラメータチューニングを行った．

評価指標として，推定した信号から音源定位により求められた音源方向（予測定位方向）と，正解の信号から音源定位により推定された音源方向（正解定位方向）の絶対誤差を DoA-A 誤差として導入した．また，推定した信号から音源定位により求められた音源方向

表 1: 各手法の IR 評価指標および音源定位誤差

(a) 実データ

モデル	Phase	Amp.	Env. [%]	T60 [%]	C50 [dB]	EDT [ms]	DoA-A [°]	DoA-B [°]
AVR	1.60±0.81	18.21±25.29	115.23±134.31	28.47±25.36	16.85±11.01	11.42±7.80	31.75±28.89	32.35±25.06
AVR+	1.58±0.82	0.77±1.39	7.51±11.83	26.63±24.99	12.81±4.75	11.89±8.01	36.31±26.81	36.86±26.01
AVR++	1.61±0.80	1.29±2.49	14.26±16.48	43.65±34.78	21.90±12.32	18.63±13.63	37.75±22.59	32.83±23.12
NAF	1.62±0.80	0.98±0.02	9.39±14.69	26.56±15.63	14.34±6.71	65.76±14.99	33.06±26.65	28.99±21.16
NAF+	1.62±0.81	0.80±0.16	8.04±13.09	30.46±17.71	13.35±6.99	31.69±14.15	**20.25±14.54**	**18.65±12.98**

(b) シミュレーション 1

モデル	Phase	Amp.	Env. [%]	T60 [%]	C50 [dB]	EDT [ms]	DoA-A [°]	DoA-B [°]
AVR	1.61±0.81	0.46±0.29	8.81±14.71	22.09±10.74	5.88±2.86	54.18±37.60	43.38±32.56	42.24±33.52
AVR+	1.62±0.80	1.61±1.90	11.22±15.60	44.83±9.73	16.55±3.64	202.46±61.66	30.13±13.43	29.14±13.77
AVR++	1.62±0.80	0.52±0.57	9.44±12.84	13.01±9.62	2.30±1.63	52.91±38.97	34.94±19.89	34.81±19.44
NAF	1.60±0.80	0.56±0.09	5.15±8.64	14.21±9.00	1.39±0.93	44.90±32.93	29.69±27.52	28.98±26.75
NAF+	1.62±0.81	0.58±0.10	4.59±8.13	9.42±7.21	1.49±1.11	36.78±29.92	**27.25±20.67**	**27.40±20.37**

(c) シミュレーション 2

モデル	Phase	Amp.	Env. [%]	T60 [%]	C50 [dB]	EDT [ms]	DoA-A [°]	DoA-B [°]
AVR	1.58±0.81	0.29±0.14	9.34±10.39	14.18±8.30	3.00±1.82	19.77±13.99	43.06±26.09	48.09±26.13
AVR+	1.52±0.82	0.46±0.11	8.30±9.79	15.54±6.36	2.92±1.60	20.15±11.02	32.31±21.46	**29.88±20.42**
AVR++	1.62±0.80	0.64±0.18	9.92±11.79	16.87±8.07	8.01±2.61	25.51±14.83	32.63±25.03	30.26±25.27
NAF	1.62±0.81	0.85±0.07	8.14±11.05	31.60±10.14	8.13±1.42	36.62±21.75	**30.19±17.09**	31.83±17.85
NAF+	1.62±0.81	0.61±0.08	8.28±11.34	18.00±11.94	1.26±0.95	26.97±17.51	33.50±28.71	30.81±26.37

表 2: ハイパーパラメータ（NAF, NAF+）

項目	範囲
層数	4 〜 12
残差層の層数	0 〜 3
中間層の特徴量数	64 〜 1024
座標に加えるノイズ量	0.0 〜 0.2
初期学習率	1×10^{-5} 〜 1×10^{-2}
学習率減衰率	0.05 〜 0.5
位相損失の重み	0.1 〜 10
1 step のサンプリング点数（時刻×周波数）	200 〜 4000

表 3: ハイパーパラメータ（AVR, AVR+，AVR++）

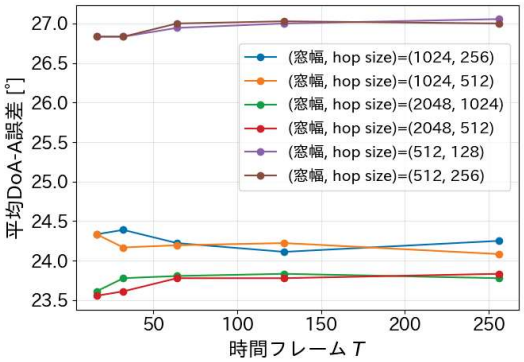
項目	範囲
初期学習率	1×10^{-6} 〜 1×10^{-4}
最小学習率	初期学習率の 1% 〜 50%
半径方向のサンプル数	40 〜 80
水平角方向の分割数	48 〜 80
重み減衰	0 〜 0.001
スペクトル誤差の重み	0 〜 100
角度誤差の重み	0 〜 100
時間誤差の重み	0 〜 100
エネルギー誤差の重み	0 〜 100
Multi-STFT 誤差の重み	0 〜 100
DAS 誤差の重み	0 〜 100
σ エンコーダの幅	32 〜 512
σ デコーダの幅	32 〜 512
signal ネットワークの幅	128 〜 1024
σ エンコーダに埋め込み	True / False
σ デコーダに埋め込み	True / False
signal ネットワークに埋め込み	True / False

（予測定位方向）と，実際の音源方向（真の音源方向）の絶対誤差を DoA-B 誤差と定めた．

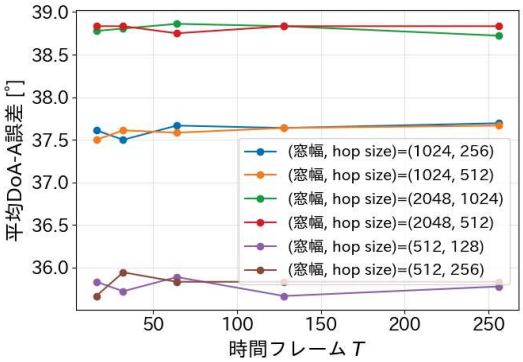
ニューラルネットではモデルを教師信号に近づけるよう学習するため，検証の評価指標としては DoA-A 誤差の平均を最小化するようにチューニングを行った．音源定位の手法には MUSIC 法を用い，STFT により得られた時間-周波数領域のインパルス応答を利用して音源定位を行った．STFT のパラメータは窓幅を 512，hop size を 128，窓関数をハニング窓とした．各試行において 50 epochs 学習を行い，各 epoch 終了時に検証データに対する DoA-A 誤差の計算を行って，その最小値を記録した．

2 節で提案した AVR，AVR+，AVR++，NAF，NAF+ に対して 100 回の試行を行った際の最良の結果を表 1a，表 1b，表 1c に示す．表 1a が実データセット，表 1b がシミュレーション 1，表 1c がシミュレーション 2 に対する結果を表す．

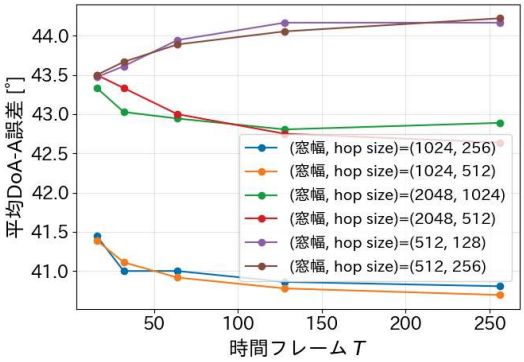
DoA-A 誤差及び DoA-B 誤差は外れ値による分散への影響が大きかったため，誤差最大値の 90 パーセンタイルを基準として外れ値除去を行った結果となっている．さらに，IR 評価指標として，位相 Phase，振幅 Amp.，包絡線 Env.，残響時間 T60，明瞭度 C50，初期残響時間 EDT に対する誤差を表に示している．表 1a，表 1b より，実データ及びシミュレーションデータ



(a) 実データ



(b) シミュレーション 1



(c) シミュレーション 2

図 6: ホワイトノイズ音源での音源定位誤差 DoA-A

1 に対しては NAF+の音源定位誤差が最小となっていることがわかる．表 1c では，DoA-A で NAF，DoA-B で AVR+が最小の誤差を示しているが，NAF+も小さい誤差となっていることがわかる．これらの結果から，音源定位誤差で比較すると，提案手法の中で NAF+が最良のモデルであるといえる．

次に，検証データに対して NAF+が推定した各インパルス応答について，音源定位による評価を行った結果を図 5a，図 5b，図 5c に示す．図 5a が実データセットについての結果，図 5b がシミュレーション 1 についての結果，図 5c がシミュレーション 2 についての結果を表す．左側の図は横軸が正解インパルス応答による音源定位方向，縦軸が推定インパルス応答による音源定位方向を表す．すなわち，誤差 DoA-A を表す．右側の図は横軸が実際の音源方向，縦軸が推定インパルス応答による音源定位方向を表す．すなわち，誤差 DoA-B を表す．いずれの図も，斜めに引かれた赤の破線上に点が近いほど誤差が小さいことを表す．なお，縦軸横軸ともに範囲が 0° から 360° であり，実際には上下と左右の端は連続していることに注意する．いずれのグラフでも，数点の外れ値を除いて，破線上に近い位置に音源定位結果が分布していることがわかる．

4.2 実験 2: ホワイトノイズ音源での音源定位評価

現実には，4.1 節で行ったようなインパルス信号で音源定位をする状況は非常に限られている．そこで，実用的な長さの音響信号で音源定位をするため，音源を 100s のホワイトノイズとしたときの音源定位評価を行った．具体的には，100s のホワイトノイズをインパルス応答と畳み込み，STFT 後に時間フレーム T ごとに区切って音源定位を行った．各時間フレームでの音源定位方向の中央値を全体での音源定位方向とし，DoA-A 誤差，DoA-B 誤差を評価した．

検証データに対する NAF+の推定インパルス応答を用いて，音源をホワイトノイズとしたときの音源定位評価を行った．まず，最適なハイパーパラメータを決めるために，時間フレーム T を 16，32，64，128，256 の 5 パターン，さらに，STFT のパラメータ（窓幅，hop size）を（512，128），（512，256），（1024，256），（1024，512），（2048，512），（2048，1024）の 6 パターン使用して，最適なパラメータの検討を行った．なお，STFT の窓関数はハニング窓を使用した．各ハイパーパラメータでの DoA-A 誤差の結果を図 6a，図 6b，図 6c にまと

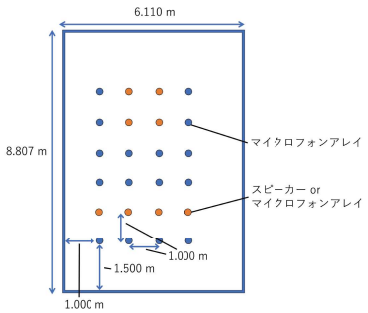


図 7: 音源位置 (図 3 より，音源位置（橙点）を変更)

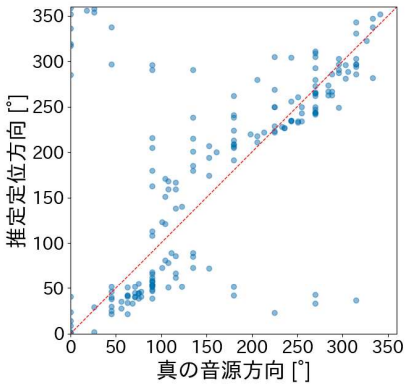


図 8: インパルス音源

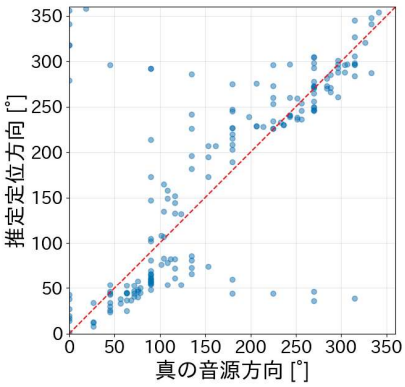


図 9: ホワイトノイズ音源

表 4: 音源による誤差の違い

(a) DoA-A 誤差

音源	実データ	シミュレーション 1	シミュレーション 2
インパルス	20.25° ± 14.54°	27.25° ± 20.67°	33.50° ± 28.71°
ホワイトノイズ	18.50° ± 16.20°	27.72° ± 20.06°	30.34° ± 26.73°

(b) DoA-B 誤差

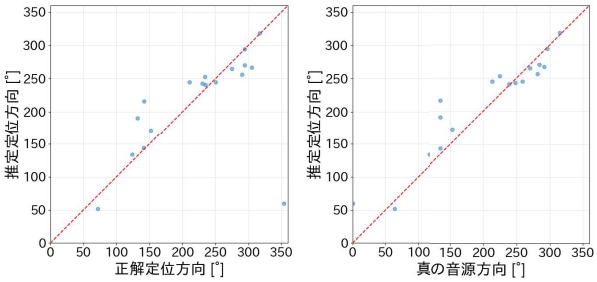
音源	実データ	シミュレーション 1	シミュレーション 2
インパルス	18.65° ± 12.98°	27.40° ± 20.37°	30.81° ± 26.37°
ホワイトノイズ	16.71° ± 14.16°	27.77° ± 19.65°	28.60° ± 26.04°

めた．STFT パラメータについては，いずれのデータセットでも窓幅 = 1024，hop size= 512 において平均的に誤差が小さくなっていることがわかる．一方，時間フレーム T の変化に対しては，あまり誤差が変化していないことがわかる．

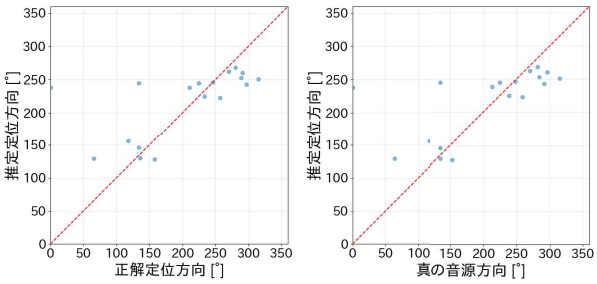
次に，時間フレーム $T = 64$，STFT パラメータ (窓幅=1024，hop size=512) としてホワイトノイズ音源での音源定位評価を行った．その結果を図 10a，図 10b，図 10c に示す．左側の図が誤差 DoA-A，右側の図が誤差 DoA-B を表す．いずれの図も，斜めに引かれた赤の破線上に点が近いほど誤差が小さいことを表す．いずれのグラフでも，数点の外れ値を除いて，破線上に近い位置に音源定位結果が分布していることがわかる．また，音源がインパルス信号の場合の音源定位誤差と比較した結果を表 4a，表 4b に示す．なお，表の値は誤差最大値の 90 パーセンタイルを基準として外れ値除去を行った結果となっている．インパルス音源のときと比較してホワイトノイズ音源では誤差がやや小さくなっていることがわかる．

4.3 実験 3: 未知の音源位置に対する音源定位評価

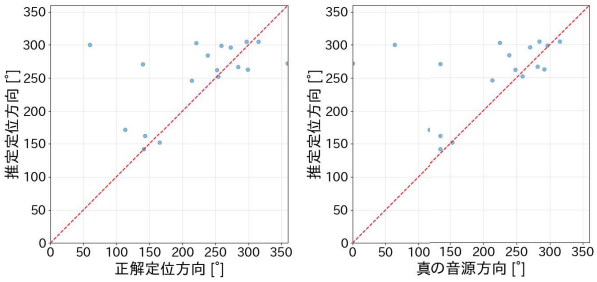
4.1 節，4.2 節において DoA-A 及び DoA-B 誤差が最小だった，実データセットで学習した NAF+について、同一環境の未知音源位置に対する音源定位評価を行った．学習データに含まれる音源位置と重ならないように，図 7 における橙色の 8 箇所をスピーカーの配置候補箇所とし，スピーカー位置を除く格子上の 23 点をマイクロフォンアレイ配置候補場所とした．なお，格子点の配置は学習データの配置図 3 と同じものを使用した．合計で $8 \times 23 = 184$ 個の音源，マイク位置ペアを作成し，学習済みの NAF+でインパルス応答の推定を行った．本実験では，正解の信号がなく DoA-A 誤差は計算できないため，インパルス音源及びホワイトノイズ音源での DoA-B 誤差の結果を，それぞれ図 8，図 9 に示す．90 パーセンタイル基準で外れ値除去すると，インパルス音源での DoA-B が 26.23° ± 17.33°，ホワイトノイズ音源では 24.00° ± 16.68° であった．このことから，未知の音源位置に対しても，検証データと同等の音源定位誤差となることが確認できた．



(a) 実データ



(b) シミュレーション 1



(c) シミュレーション 2

図 10: ホワイトノイズ信号での音源定位分布

4.4 おわりに

本研究では，多チャンネル音場推定手法の評価指標として，マイクロフォンアレイによる音源定位誤差に着目した検討を行った．実環境で計測したインパルス応答データセットに加え，AcoustiX および Pyroomacoustics により構築したシミュレーションデータセットを用いて，AVR，NAF 及びそれらを多チャンネルデータに拡張した AVR+，AVR++，NAF+の比較を行った．その結果，提案したチャンネル埋め込みを導入した NAF+が最も小さい音源定位誤差を示すことが確認された．また，ホワイトノイズを音源とする実用的な長さの音響信号を用いて音源定位評価を行ったところ，インパルス音源の場合と比較して誤差がやや小さくなることを確認した．さらに，実データセットで学習した NAF+を用いて未知の音源位置に対する音源定位評価を行い，検証データと同程度の誤差で音源方向推定が可能であることを確認した．

今後の課題として，より多いデータ数や多様な音響環境で提案手法を学習させることで，推定信号からの音源定位誤差をさらに小さくすることなどが考えられる．

Acknowledgements

本研究は CREST JPMJCR22D3 および RIKEN TRIP initiative (AGIS) の助成を受けた．

参考文献

[Heath 24] Heath, B. E., Suzuki, R., Le Penru, N. P., Skinner, J., Orme, C. D. L., Ewers, R. M., Sethi, S. S., and Picinali, L.: Spatial Ecosystem Monitoring with a Multichannel Acoustic Autonomous Recording Unit (MAARU), *Methods in Ecology and Evolution*, Vol. 15, No. 9, pp. 1568–1579 (2024)

[Hoydis 22] Hoydis, J., Cammerer, S., Ait Aoudia, F., Nimier-David, M., Maggi, L., Marcus, G., Vem, A., and Keller, A.: Sionna (2022), https://nvlabs.github.io/sionna/

[Lan 24] Lan, Z., Zheng, C., Zheng, Z., and Zhao, M.: Acoustic Volume Rendering for Neural Impulse Response Fields, *arXiv preprint arXiv:2411.06307* (2024)

[Luo 22] Luo, A., Du, Y., Tarr, M., Tenenbaum, J., Torralba, A., and Gan, C.: Learning neural acoustic fields, *Advances in Neural Information Processing Systems*, Vol. 35, pp. 3165–3177 (2022)

[Mildenhall 20] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, in *ECCV* (2020)

[Mošner 24] Mošner, L., Serizel, R., Burget, L., Plchot, O., Vincent, E., Peng, J., and Černocký, J.: Multi-Channel Extension of Pre-trained Models for Speaker Verification, in *Interspeech 2024* (2024)

[Nakadai 20] Nakadai, K. and Okuno, H. G.: Robot Audition and Computational Auditory Scene Analysis, *Advanced Intelligent Systems*, Vol. 2, No. 9, p. 2000050 (2020)

[Salvati 14] Salvati, D., Drioli, C., and Foresti, G. L.: Incoherent Frequency Fusion for Broadband Steered Response Power Algorithms in Noisy Environments, in *IEEE Signal Process. Lett.*, Vol. 21, pp. 581–585 (2014)

[Scheibler 18] Scheibler, R., Bezzam, E., and Dokmanić, I.: Pyroomacoustics: A Python package for audio room simulations and array processing algorithms, in *IEEE ICASSP* (2018)

[加藤 24] 加藤 雅大, 小島 諒介：多チャンネル音響生成を評価するためのマイクロフォンアレイ室内インパルス応答データセットの構築, 人工知能学会 第 128 回人工知能基本問題研究会（SIG-FPAI） (2024)

[平塚 24] 平塚 謙良, 小島 諒介：世界モデルベースの深層強化学習による音源追跡の検討, 第 66 回人工知能学会 AI チャレンジ研究会 (2024)

[平塚 25] 平塚 謙良, 小島 諒介：音環境情報に基づくピックアンドプレースの模倣学習, RSJ2025 (2025)

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-068-04

OWSM v3.1を活用した 既知・未知混合条件における話者ダイアライゼーション

Speaker Diarization in Mixed Known-Unknown Conditions utilizing OWSM v3.1

阿坂 脩平 [1*]　Muhammad Shakeel[2]　前田 力 [2]　Benjamin Yen[1]
芦澤 剛 [1]　住田 直亮 [2]　中臺 一博 [1]

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

Abstract: 本稿では，実運用を想定し，既知話者と未知話者が混在する条件における話者ダイアライゼーションの性能向上を目的とする．従来手法では部分的な既知話者情報を活用できないため，本研究では EEND-VC に基づき，以下の 2 点を提案する：1) セグメンテーション性能を底上げする OWSM-Encoder の導入，2) 既知情報を反映する既知シード誘導クラスタリングの導入．実会議コーパスを用いた評価により，OWSM-Encoder が従来モデルと比較して競争力のある結果を示すこと，さらに既知シード誘導クラスタリングの導入によって最大 4.17 ポイントの DER の改善を確認した．

1　はじめに

話者ダイアライゼーションとは,「いつ」「誰が」発話したかを複数話者の会話音声から推定する技術である．高性能な話者ダイアライゼーションを実現することで，その結果を音声認識の補助情報として利用し，高品質な文字起こしを可能にできる．応用例としては，企業等における自動会議録作成が挙げられる．一般的な会議を想定すると，内部参加者については事前に登録された音声特徴を手掛かりとして話者ダイアライゼーションを行える一方，外部参加者の音声特徴は未知であることが多い．話者ダイアライゼーション手法として，ターゲット話者検出に基づく手法 [1, 2] は話者特徴を活用可能だが，全話者の特徴を必要とする．また，近年主流となっている EEND-VC 系列の手法 [3, 4] は事前の話者情報がないことを前提としており，既知情報を活用できない．したがって，既知話者と未知話者が混在する状況で，既知話者の情報を活用できる話者ダイアライゼーション手法の検討が求められる．

本稿では，EEND-VC アーキテクチャを基盤として，部分的な既知情報を活用可能な話者ダイアライゼーションシステムを構築する．具体的には，セグメンテーション部に音声認識タスク向けに事前学習された OWSM [5] のエンコーダを導入し，音響情報に加えて言語的な情報も利用することで，発話区間検知の性能向上を図る．さらに，グローバルクラスタリング部では既知話者の埋め込みを用いた既知シード誘導クラスタリングを導入し，部分的に得られる既知話者情報を活用してクラスタリング性能を向上させる．

評価実験には AMI [6]，AliMeeting [7]，AISHELL-4 [8] の 3 つの遠距離単一チャネル音声コーパスを用いた．その結果，OWSM-Encoder の導入により DER の改善が確認され，他のモデルと比較して競争力のある結果が得られた．また，既知シード誘導クラスタリングにより部分的な既知話者情報を活用することで，最大 4.17 ポイントの DER の改善を確認した．

2　関連研究

話者ダイアライゼーションの従来手法を，既知話者手法と未知話者手法の二つに大別し，それぞれが既知・未知混合条件に有効に対処できない点を指摘する．本稿では，登場話者全員の音声特徴を登録音声として事前に用意し，それを活用して話者ダイアライゼーションを行う手法を既知話者手法とする．一方で，話者数や話者特徴量に関する事前情報を用いず，複数人対話音声のみから発話区間検知を行う手法を未知話者手法と定義する．

*連絡先： 東京科学大学
152-8552 東京都目黒区大岡山 2-12-1
E-mail: asaka@ra.sc.e.titech.ac.jp

2.1 既知話者手法

既知話者手法としては，Ding らの Personal VAD [1] や，Medennikov らの Target-Speaker VAD（TS-VAD）[2] が挙げられる．Personal VAD は，事前に取得した音声特徴量を基に，特定話者の発話をフレーム単位で判定する手法である．この考え方は，特定話者を対象とした音声認識 [9] にも応用されている．しかし，一度に一人の話者しか扱えないという制約がある．一方 TS-VAD は，4 人分の音声特徴量を用いて同時に発話区間検知を行い，重複発話にも頑健なダイアライゼーションを実現した．ただし，推論には登場話者全員の音声特徴量が必要である．さらに， TS-VAD 自体は話者特徴推定を含む未知話者手法の研究を目的としているため，既知話者手法としての具体的評価は行われていない．これは，登録音声を利用できる複数話者の実環境コーパスが不足しているためと考えられる．以上より，従来の既知話者手法は未知話者混在条件に対応できず，実環境コーパスを用いた検証も十分ではない．

2.2 未知話者手法

未知話者手法としては，End-to-End Neural Diarization (EEND) [10] が深層学習を活用した手法として有力視された．EEND は同時発話に対応可能な出力形式を備える一方，長時間音声や話者数変動への対応には課題があった．その後，EEND とクラスタリングベースの手法を統合した EEND-VC [3, 4] が木下らによって提案された．この手法では，まずセグメンテーション部において短い音声チャンクごとに EEND を適用し，複数人の重複を含む発話区間を検知する．次に，グローバルクラスタリング部において，その推定発話区間から話者埋め込みを取得する．その後，凝縮型クラスタリングによって推定話者ラベルを付与することで，長時間音声への対応と可変話者数の処理を可能にしている．このアーキテクチャは pyannote フレームワーク [11] を通じて広く利用されており，近年では自己教師あり学習を用いた手法の検討 [12, 13]，音源分離との統合 [14]，さらにクラスタリング手法の改善 [15] などが進められている．しかし，あくまで未知話者手法であるため，本稿が対象とする既知・未知混合条件において，既知話者情報を有効に活用することはできない．

3 提案手法

既知・未知混合条件において部分的な情報を有効に活用できる話者ダイアライゼーションを実現するため，本稿では前節で述べた有力な未知話者手法である EEND-VC を基盤として，二つの提案手法を導入した話者ダイアライゼーションシステムを構築する．提案手法は以下の二点である．第一に，セグメンテーション性能向上のために OWSM v3.1 [5] のエンコーダ部をセグメンテーションモデルへ導入し，言語的な内容に基づく発話区間推定を可能にする．第二に，部分的に得られる既知話者情報を活用するため，既知シード誘導クラスタリングを導入する．本稿では，DiariZen [1] が提供する EEND-VC アーキテクチャを基に，図 1 に示す提案アーキテクチャを構築した．

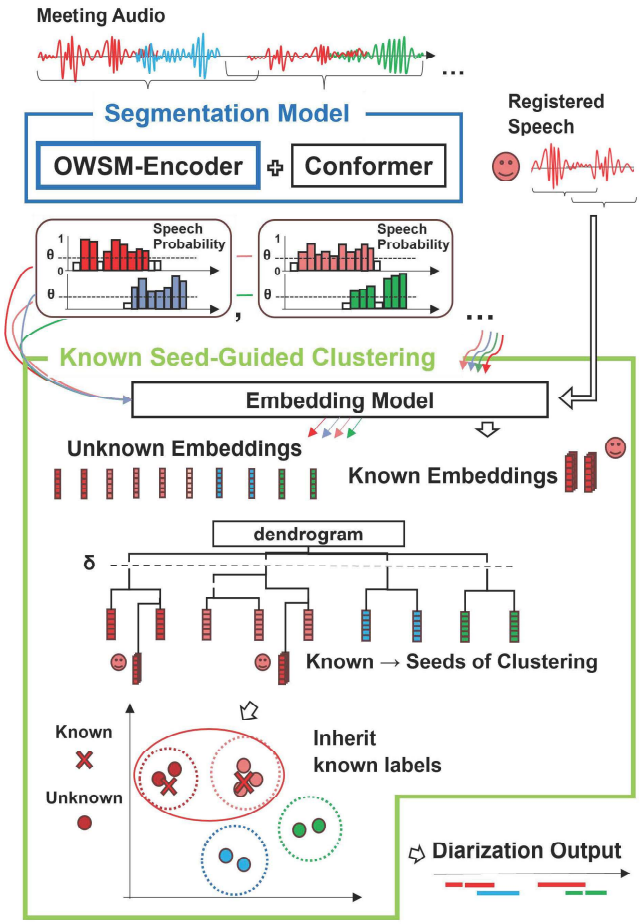


図 1: Proposed Speaker Diarization System

3.1 OWSM-Encoder の導入

既知話者情報の活用効果はセグメンテーション品質に大きく左右されるため，本研究ではまずセグメンテーション性能の向上を目的として OWSM-Encoder を導入する．OWSM（Open Whisper Style Speech Model）[16] は，Whisper [17] と同様の学習をオープンソースツールと公開データのみで再現することを目的としたモデルである．大規模音声データを用いて音声認識およ

[1] https://github.com/BUTSpeechFIT/DiariZen

び翻訳を学習していることから,「汎用音声基盤モデル」としての応用が期待され，実際に幅広い音声タスクで利用されている [18]．参考としたセグメンテーションモデルで用いられている WavLM [12] は，自己教師あり学習を通じて音響的な文脈情報を獲得している一方，言語的・意味的な文脈を明示的に学習しているわけではない．そこで，音声–言語タスクに特化した OWSM に置き換えることで，発話内容に基づく言語的手がかりを活用した発話区間検知が可能になると考えられる．加えて，WavLM と異なり OWSM は事前学習コードが公開されているため，再現性やカスタマイズ性が高い点も利点である．

本稿では E-branchformer [19] を導入することで性能向上が確認された OWSM v3.1 [5] を採用する．導入にあたって注意すべき点として，WavLM は音声波形入力であるのに対し，OWSM のエンコーダ部はメルスペクトログラムを入力とする．そのため，OWSM の入力層で用いられている STFT を併用し，特徴量抽出を担う E-branchformer Encoder とまとめて OWSM-Encoder として導入する．最終的に，セグメンテーションモデルでは OWSM-Encoder による特徴量抽出を行い，その最終層出力を Conformer [20] に入力することで，チャンク単位の推定話者発話区間検知を行う．ただし，音声認識・翻訳向けに学習されたモデルをそのままセグメンテーションに用いることは難しいため，本研究では LibriMix [21] を用い，ESPnet [22] により提供されている話者ダイアライゼーションタスクでファインチューニングを行った．

3.2 既知シード誘導クラスタリング

部分的に得られる既知話者情報を活用して話者ダイアライゼーション性能を向上させるため，既知シード誘導クラスタリングを導入する．従来の EEND-VC における凝縮型クラスタリングは以下の手順で行われる：1) 特徴量抽出器を用いて各チャンクの推定話者の特徴ベクトルを全区間から取得し正規化する，2) ユークリッド距離に基づき樹形図を作成する，3) 距離閾値によりクラスタを生成する，4) 小規模クラスタを最近傍のクラスタに統合する．本手法では，このクラスタリング過程に既知話者特徴ベクトルを組み込む．具体的には，各既知話者の音声のうち指定秒数を抽出し，1) と同じ特徴量抽出器に入力して正規化済みの埋め込みベクトル（既知話者特徴ベクトル）を生成する．既知話者一人あたり複数の特徴ベクトルが得られるが，これらを複製して，同一の特徴量セットを作成する．次に，通常の特徴ベクトルに加えて，既知話者特徴ベクトルも含めた形で樹形図を作成する．この操作により，複製した既知話者特徴ベクトルがクラスタ初期段階でまとまり，クラスタリングのシードとして機能する．さらに，複製倍率が重みとして働き，クラスタの重心を既知話者特徴ベクトル側へ引き寄せる効果が生じ，既知話者情報を基軸としたクラスタリングが可能となる．クラスタ生成後は，既知話者特徴ベクトルを含むクラスタのラベルをその話者に変更し，誤った分割が生じていても同一話者クラスタとして修正可能とする．もし複数の既知話者が同一クラスタに含まれていた場合は，含まれる既知話者特徴ベクトル数が最も多い話者をクラスタラベルとして採用する．このように，従来の凝縮型クラスタリングの枠組みを大きく変更することなく既知話者特徴ベクトルを導入することで，それによるエラーを防ぎつつ，部分的な既知話者情報を活用できる．

4 評価実験

評価実験は二つの観点から行った．実験 1 では OWSM-Encoder 導入による性能向上を評価し，実験 2 では既知シード誘導クラスタリングによる性能向上を評価する．

4.1 データセット

評価には DiariZen の設定と同様に，AMI (AMI-SDM) [6]，AliMeeting [7]，AISHELL-4 [8] の遠距離単一チャンネル音声を用いた．AISHELL-4 のみ，各チャンネル音声を平均した 1 チャンネル音声を使用し，検証データはトレーニングデータの 10% を抽出して作成した．学習・検証データおよびデフォルトの評価データの詳細を表 1 に示す．また，既存の評価データはそのままでは既知・未知混合条件の評価ができないため，構造を再構成したデータを表 2 の通りに用意した．図 2 に示すように，各音声ファイルの冒頭から重複のない発話区間を切り出し，その話者の登録音声として話者ラベルと紐付けて保存した．各ファイルにつき，登場話者 3 名分について，累計 20 秒ずつの登録音声を作成し，その時点で音声を切り分け，残りを評価データとした．AliMeeting については話者数が 3 名未満のファイルを除外し，AISHELL-4 については音声切り出しによって，特定話者の発話が評価データから消失したファイルを除外した．これらの切り出し・除外処理による評価データの縮小率を表 2 に示す．学習と検証には三つのデータセットを併用し，評価は各データセットごとに行う．

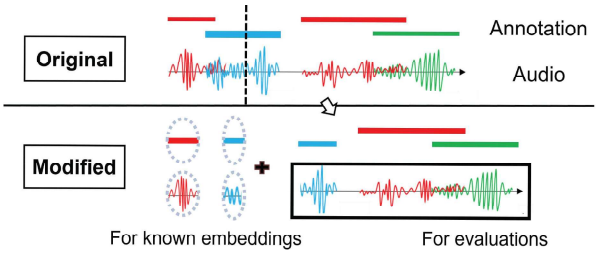


図 2: Method for Creating Evaluation Data

表 1: Dataset

Dataset	Train			Dev			Test		
	files	spk	hrs	files	spk	hrs	files	spk	hrs
AMI	134	3-5	80	18	4	10	16	3-4	9
AliMeeting	209	2-4	111	8	2-4	4	20	2-4	11
AISHELL-4	173	3-7	97	18	3-7	10	20	5-7	13
whole	516	2-7	288	44	2-7	24	56	2-7	33

4.2 学習・推論設定

OWSM-Encoder のファインチューニングには ESPnet [2] を用い，ダイアライゼーションモデルには EEND-EDA [23] を使用した．入力 STFT の設定は FFT 512，窓サイズ 400，フレームシフト 160 とした．学習率は 4e-4，warmup step は 100,000，最適化は AdamW [24] を用い，バッチサイズは 20 とした．データセットは LibriMix [21] を用い，Libri2Mix を 30 epoch，Libri3Mix を 10 epoch 学習し，best モデルを採用した．

セグメンテーションモデルの学習には DiariZen [3] を使用した．比較対象として，ファインチューニングした OWSM v3.1 medium のほか WavLM-base+，WavLM-large を用い，いずれも最終層出力を線形層に通して Conformer に入力した．学習時には事前学習済みモデルも更新対象とし，チャンク 8 秒，ストライド 3 秒，損失関数は Powerset loss [25]，最大話者数は 4 とした．学習率は WavLM／OWSM 部分を 1e-5，その他を 1e-3 とし，実効バッチサイズ 64，最適化は AdamW を使用した．最大 100 epoch 学習し，10 epoch 改善がない場合は早期終了した．AutoClip [26] により，勾配ノルムの 90 パーセンタイルに基づきクリッピング閾値を設定した．学習は TSUBAME4.0 node-q（NVIDIA H100 SXM5 94GB）で実施した．

推論時には，事前話者音声を扱えるよう改変した pyannote pipeline [11] を使用し，セグメンテーションモデルは損失が小さい 5 エポックの平均を用いた．推論設定は，チャンク 8 秒，ストライド 0.8 秒，話者数 1〜20，人クラスタリング閾値 0.7，最小クラスタサイズ 30 とした．特徴量抽出には Wespeaker [27, 28] の

[2] https://github.com/espnet/espnet
[3] https://github.com/BUTSpeechFIT/DiariZen

表 2: Modified Test Dataset

Dataset	Modified Test					
	files	known spk	known sec	spk	test hrs	retained rate
AMI	16	3	20	3-4	6.9	76%
AliMeeting	12	3	20	3-4	5.8	53%
AISHELL-4	19	3	20	5-7	9.7	76%
whole	56	3	20	3-7	22.4	69%

ResNet34-LM [4] を使用した．

既知話者特徴ベクトルは，一人当たり 20 秒の登録音声を用い，以下の 3 パラメータで切り出した音声片を特徴量抽出器に通して取得した：

- **Size**: 切り出す音声片の長さ
- **Step**: 音声片を切り出す間隔
- **Replication Factor**: 音声片の複製倍率

本稿では Size = 5.0s，Step = 0.8s を採用した．Size は，チャンク 8 秒内で話者の一貫した発話が概ね 2〜5 秒であること，および既存モデルが 2〜3 秒で学習されていること [29, 30] に基づき，より安定した表現を得るために決定した．Step は，埋め込みの整合性を保つためセグメンテーションのストライドと揃えた．Replication Factor は適切な指標がないため，クラスタリングへの寄与を実験的に探索した．

4.3 評価指標

話者ダイアライゼーションの評価には Diarization Error Rate (DER) を用い，次式のように定義する：

$$\mathrm{DER} = \frac{\text{false alarm}+\text{mis-detection}+\text{confusion}}{\text{total}} \tag{1}$$

ここで，**false alarm** は非発話を誤って発話とした区間，**mis-detection** は発話を誤って非発話とした区間，**confusion** は発話区間における話者誤識別の区間，**total** は全話者の正解発話区間の合計時間である．

本稿では，話者交代時の誤差許容（collar）を設けず，最も厳しい条件で DER を算出する [31]．DER の算出には pyannote-metric [32] を用いる．

4.4 実験 1：OWSM-Encoder の性能比較

OWSM-Encoder の導入によるセグメンテーション性能向上の寄与を評価する．比較手法として WavLM-base+ と WavLM-large を用い，最先端手法との比較のため，未知話者のみの条件であるデフォルトの評価データを使用した．

[4] https://huggingface.co/pyannote/wespeaker-voxceleb-resnet34-LM

4.5 結果 1：OWSM-Encoder の性能比較

比較手法の DER を表 3 に示す．表 3 より，AMI と AliMeeting では WavLM-large には及ばないものの，WavLM-base+ と比較すると競争力のある結果が得られた．AMI において性能向上が得られなかった理由としては，ノイズが大きく環境音が多様であるため，言語的手がかりが十分に作用しなかった可能性が考えられる．一方で AISHELL-4 では，WavLM-base+ より 0.42 ポイント，WavLM-large より 0.31 ポイント高い性能を示した．この結果から，OWSM-Encoder による言語的・意味的文脈の反映は，対象とする音声によっては有効に働くことが示された．

4.6 実験 2：提案クラスタリングの性能比較

既知シード誘導クラスタリングの導入による性能向上を評価する．未知・既知混合条件を扱うため，4.1 節で述べた，構造を変更したデータセットを用いる．まず，ベースラインとして WavLM-base+ を用い，Replication Factor を実験的に探索する．パラメータの候補は [5,10,20,50,100,200,400,750] とし，それぞれ既知話者が 1 人の場合と 3 人の場合について DER を算出した．次に，その結果に基づき，既知話者特徴ベクトルの抽出パラメータを確定し，WavLM-base+，WavLM-large，OWSM-Encoder を用いた場合の，既知話者人数による DER の比較を行った．既知話者の組み合わせが複数ある場合は全通りを実行し，平均値を算出した．また，2 節の通り，既知・未知混合条件で比較可能な従来手法が存在しないため，既知話者情報を利用しない場合をベースラインとする．

4.7 結果 2：提案クラスタリングの性能比較

Replication Factor を変化させたときの DER を表 4 に示す．これより，既知話者人数に依らず，Modified AMI では Replication Factor = 5〜10 が適しており，他の二つのコーパスでは Replication Factor = 100〜200 が適していることが読み取れる．したがって，既知話者特徴ベクトル抽出パラメータを (Size, Step, Replication Factor) = (5, 0.8, 10) および (5, 0.8, 200) の 2 種類に決定した．

これらのパラメータを用いた場合の既知話者人数ごとの DER を表 5 に示す．結果より，全体として既知話者情報を適用した方が，適用しない場合より良い性能を示す傾向が確認できる．しかし，既知話者人数が多いほど性能が向上するとは一概には言えず，既知話者特徴ベクトル間で分布が重なりエラーを生むケースがあると考えられる．個別のデータセットを見ると，AMI では二つのパラメータに共通して，WavLM-large の 1 既知話者適用時が最高性能であり，未適用時と比較して 0.26 ポイント改善した．AliMeeting では，Replication Factor = 200 かつ WavLM-base+ の 3 既知話者適用時が最高性能で，未適用時より 4.17 ポイント改善した．ここでは，Replication Factor が大きい，つまりは既知話者特徴ベクトルの重みが大きい方が，AliMeeting 全体として既知話者適用時の改善幅が大きい傾向が示された．これは，AliMeeting が中国語による激しいインタラクションを特徴とする複雑なコーパスであり，既知話者情報という明確な基準が，話者識別において有効に作用したためと考えられる．また，AISHELL-4 では，Replication Factor = 200 かつ WavLM-base+ の 2 既知話者適用時が最高性能で，未適用時より 0.12 ポイント改善した．ただし，既知話者人数によっては改善と同程度のエラーも発生している．これらの結果より，コーパスの特徴に応じて，既知シード誘導クラスタリングの効果が大きく左右されることがわかった．

(a) Without known information

(b) 3 Known Speakers Applied

図 3: Clusters of R8002_M8003_MS803 in AliMeeting

さらに，既知シード誘導クラスタリングの効果を視覚的に確認するため，UMAP [35] による可視化を行った．図 3 には，性能向上が大きかった AliMeeting における最終的なクラスタ分布を示す．これより，既知話者を適用することで，従来は混同していた話者クラスタが，既知話者特徴ベクトルを手掛かりとして分離できていることが確認され，提案手法の有効性が裏付けられた．

表 3: Performance Comparison of OWSM-Encoder

Model	AMI				AliMeeting				AISHELL-4			
	DER	FA	MD	CN	DER	FA	MD	CN	DER	FA	MD	CN
WavLM-base+	15.54	3.69	8.26	3.59	18.26	2.52	8.41	7.33	10.67	4.10	3.57	3.00
WavLM-large	**14.55**	3.76	7.40	3.39	**15.11**	2.46	8.00	4.65	10.56	4.81	2.69	3.06
OWSM-Encoder	16.06	3.53	8.99	3.54	16.95	2.52	8.44	5.99	**10.25**	4.59	2.58	3.08
SOTA	14.0 [33]				12.5 [34]				9.8 [33]			

表 4: Experimental Exploration of Replication Factor

	Replication Factor								
num_known = 1	0	5	10	20	50	100	200	400	750
Modified AMI	16.58	**16.56**	16.57	16.61	16.61	16.61	16.62	16.62	16.62
Modified AliMeeting	24.75	24.80	23.24	23.25	23.23	23.04	**22.44**	22.61	22.61
Modified AISHELL-4	12.34	12.33	12.38	12.31	**12.24**	**12.24**	12.25	**12.24**	12.31
	Replication Factor								
num_known = 3	0	5	10	20	50	100	200	400	750
Modified AMI	16.58	**16.49**	**16.49**	16.64	16.65	16.65	16.65	16.65	16.66
Modified AliMeeting	24.75	22.43	22.44	22.42	22.41	20.58	20.58	20.59	**20.57**
Modified AISHELL-4	12.34	12.31	12.26	**12.25**	12.26	**12.25**	12.26	12.26	12.47

5 おわりに

本稿では，既知話者と未知話者が混在する状況で，部分的な情報を活用できる話者ダイアライゼーションを実現することを目標とした．そのために，EEND-VC を基に，セグメンテーション性能向上のための言語的文脈を扱う OWSM-Encoder の導入と，部分的既知話者情報を活用出来る既知シード誘導クラスタリングの導入を行ったアーキテクチャを提案した．実会議コーパスを用いて，提案手法の評価を行ったところ，OWSM-Encoder は従来モデルと競争力のある結果を示し，既知シード誘導クラスタリングの導入によって，DER の改善を確認した．今後は，コーパスに性能を左右されない手法の提案と，セグメンテーション部への既知情報の活用を目指す．

参考文献

[1] Ding, S. et al.: Personal VAD: Speaker-Conditioned Voice Activity Detection, *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, pp. 433–439 (2020)

[2] Medennikov, I. et al.: Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario, *INTERSPEECH*, pp. 274–278 (2020)

[3] Kinoshita, K. et al.: Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds, *ICASSP*, pp. 7198–7202 (2021)

[4] Kinoshita, K. et al.: Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech, *arXiv preprint arXiv:2105.09040* (2021)

[5] Peng, Y. et al.: OWSM v3.1: Better and faster open whisper-style speech models based on e-branchformer, *arXiv preprint arXiv:2401.16658* (2024)

[6] Carletta, J. et al.: The AMI meeting corpus: A pre-announcement, *International Workshop on Machine Learning for Multimodal Interaction*, pp. 28–39 (2006)

[7] Yu, F. et al.: M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge, *ICASSP*, pp. 6167–6171 (2022)

[8] Fu, Y. et al.: Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario, *arXiv preprint arXiv:2104.03603* (2021)

[9] Maeda, C. *et al.*: Joint Target-Speaker ASR and Activity Detection, *Proc. INTERSPEECH*, pp. 1683–1687 (2025)

[10] Fujita, Y. et al.: End-to-end neural speaker diarization with self-attention, *IEEE ASRU*, pp. 296–303 (2019)

[11] Bredin, H.: pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe, *INTERSPEECH*, pp. 1983–1987 (2023)

[12] Chen, S. et al.: Wavlm: Large-scale self-supervised pre-training for full stack speech processing, *IEEE JSTSP*, Vol. 16, No. 6, pp. 1505–1518 (2022)

[13] Han, J. et al.: Leveraging self-supervised learning for speaker diarization, *ICASSP*, pp. 1–5 (2025)

表 5: Performance Comparison of Proposal Clustering Methods

With Params: (Size, Step, Replication Factor) = (5, 0.8, 10)

Model	Modified AMI				Modified AliMeeting				Modified AISHELL-4			
num_known	0	1	2	3	0	1	2	3	0	1	2	3
WavLM-base+	16.58	16.57	16.52	16.49	24.75	23.24	22.45	22.44	12.34	12.38	12.35	**12.26**
WavLM-large	16.17	**15.91**	15.93	15.96	**21.73**	22.27	22.08	22.67	12.35	12.62	12.48	12.28
OWSM-Encoder	17.26	17.35	17.37	17.38	22.67	22.89	22.71	23.25	12.28	12.45	12.41	12.29

With Params: (Size, Step, Replication Factor) = (5, 0.8, 200)

Model	Modified AMI				Modified AliMeeting				Modified AISHELL-4			
num_known	0	1	2	3	0	1	2	3	0	1	2	3
WavLM-base+	16.58	16.62	16.62	16.65	24.75	22.44	21.03	**20.58**	12.34	12.25	**12.22**	12.26
WavLM-large	16.17	**15.91**	15.92	15.94	21.73	21.43	20.60	20.74	12.35	12.52	12.41	12.37
OWSM-Encoder	17.26	17.36	17.37	17.37	22.67	22.47	21.92	22.08	12.28	12.30	12.26	12.29

[14] Kalda, J. et al.: PixIT: Joint training of speaker diarization and speech separation from real-world multi-speaker recordings, *Odyssey*, pp. 115–122 (2024)

[15] Delcroix, M. et al.: Multi-stream extension of variational bayesian HMM clustering (MS-VBx) for combined end-to-end and vector clustering-based diarization, *arXiv preprint arXiv:2305.13580* (2023)

[16] Peng, Y. et al.: Reproducing whisper-style training using an open-source toolkit and publicly available data, *IEEE ASRU Workshop*, pp. 1–8 (2023)

[17] Radford, A. et al.: Robust speech recognition via large-scale weak supervision, *PMLR International Conference on Machine Learning*, pp. 28492–28518 (2023)

[18] Shakeel, M. *et al.*: Unifying Diarization, Separation, and ASR with Multi-Speaker Encoder, *Proc. ASRU* (2025)

[19] Kim, K. et al.: E-branchformer: Branchformer with enhanced merging for speech recognition, *IEEE SLT Workshop*, pp. 84–91 (2023)

[20] Gulati, A. et al.: Conformer: Convolution-augmented transformer for speech recognition, *arXiv preprint arXiv:2005.08100* (2020)

[21] Cosentino, J. et al.: LibriMix: An Open-Source Dataset for Generalizable Speech Separation, *arXiv preprint arXiv:2005.11262* (2020)

[22] Watanabe, S. et al.: ESPnet: End-to-End Speech Processing Toolkit, *arXiv preprint arXiv:1804.00015* (2018)

[23] Horiguchi, S. et al.: End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors, *arXiv preprint arXiv:2005.09921* (2020)

[24] Loshchilov, I. and Hutter, F.: Decoupled weight decay regularization, *arXiv preprint arXiv:1711.05101* (2017)

[25] Plaquet, A. et al.: Powerset multi-class cross entropy loss for neural speaker diarization, *INTERSPEECH*, pp. 3222–3226 (2023)

[26] Seetharaman, P. et al.: Autoclip: Adaptive gradient clipping for source separation networks, *IEEE MLSP Workshop*, pp. 1–6 (2020)

[27] Wang, H. et al.: Wespeaker: A research and production oriented speaker embedding learning toolkit, *ICASSP*, pp. 1–5 (2023)

[28] Wang, S. et al.: Advancing speaker embedding learning: Wespeaker toolkit for research and production, *Speech Communication*, Vol. 162, pp. 103104 (2024)

[29] Desplanques, B. *et al.*: ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN-Based Speaker Verification, *arXiv preprint arXiv:2005.07143* (2020)

[30] Snyder, D. *et al.*: X-Vectors: Robust DNN Embeddings for Speaker Recognition, *Proc. ICASSP*, pp. 5329–5333 (2018)

[31] Landini, F. et al.: Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks, *Computer Speech & Language*, Vol. 71, p. 101254 (2022)

[32] Bredin, H.: pyannote. metrics: A Toolkit for Reproducible Evaluation, Diagnostic, and Error Analysis of Speaker Diarization Systems, *INTERSPEECH*, pp. 3587–3591 (2017)

[33] Han, J. et al.: Efficient and Generalizable Speaker Diarization via Structured Pruning of Self-Supervised Models, *arXiv preprint arXiv:2506.18623* (2025)

[34] Plaquet, A. *et al.*: Dissecting the Segmentation Model of End-to-End Diarization with Vector Clustering, *arXiv preprint arXiv:2506.11605* (2025)

[35] McInnes, L. et al.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction *arXiv preprint arXiv:1802.03426* (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-068-05

空間音響処理における物理法則に基づく機械学習

Physics-Informed Machine Learning in Spatial Audio Processing

小山 翔一 [1*]
Shoichi Koyama[1]

[1] 国立情報学研究所
[1] National Institute of Informatics

概要

音空間の解析や制御を目的とした空間音響処理技術は，その多くが複数のマイクロフォンを用いた音場の再構成／推定を基盤とし，指向性制御，バーチャルリアリティ音響，イメージング，騒音制御などに応用されてきた．従来は物理音響モデルと信号処理に基づく手法が用いられてきたが，近年は機械学習，特にニューラルネットワークを用いたデータ駆動型の手法が注目されている．このような手法は，十分な学習データが利用可能な場合には高い効果を発揮するものの，空間音響処理においては十分なデータを用意することが困難な状況は多い．このような問題は物理現象を対象とした逆問題全般において共通しており，その中で物理的な特性を機械学習技術に組み入れる，Physics-Informed Machine Learning (PIML) と呼ばれる枠組みが注目されている．本講演では，空間音響処理における PIML について概観し，特に推定対象が従う支配方程式，すなわち波動方程式あるいは Helmholtz 方程式の解空間への制約を考慮した機械学習技術について，これまでの発展と最近の取り組みを紹介する．

*連絡先： 国立情報学研究所
〒 168-0081 東京都千代田区一ツ橋 2-1-2
E-mail: koyama.shoichi@ieee.org

人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-068-06

ドローン聴覚のための計測機器搭載用治具の実飛行条件下での評価

Evaluation of Developed Structures for Drone Audition-purposed Measurement Equipment in Actual Flight Conditions

塚本悠太 [1*]　干場功太郎 [1]
Yuta Tsukamoto[1]　Kotaro Hoshiba[1]

[1] 東京科学大学
[1] Institute of Science Tokyo

Abstract: ドローンを用いた音情報による情景分析技術はドローン聴覚と呼ばれ，災害地における要救助者の探索手法として期待されている．この技術では，ドローンに計測機器を搭載するための治具が必要となる．以前の研究では，ドローンのランディングギアに取り付けることができる汎用性の高い治具を提案し，質量，耐荷重，振動特性，調整の容易さの 4 つの指標について評価した．ここで振動特性の評価は理論的な評価に限られており，実際のドローンの飛行中の評価が必要である．そこで本研究では従来の治具と提案する治具について，ドローンの飛行中の数箇所での加速度応答を比較し，治具の振動特性を評価する．その結果提案する治具は，従来の治具に比べてドローンの飛行中に計測機器に作用する振動を抑制できることが明らかになった．さらに提案する治具では，従来の治具に比べてドローンの着陸時に大きな振動が生じることが確認されたが，着陸時には音情報を収録しないため，着陸時の衝撃による収録音への影響は限定的であると考えられる．一方で提案する治具は，着陸時の衝撃に耐えるよう十分な強度を有する必要があることも明らかになった．

1　緒言

ドローンは迅速性と機動性を有しており，災害地における要救助者の探索・救助活動への適用が期待されている．そこでドローンに搭載したカメラで得られた視覚情報を利用した手法が多く検討されてきた [1–5]．しかしながら，このような視覚情報に基づく手法は，対象とする環境が暗い場合や瓦礫に埋もれた被災者を探索する場合には効果的でないという課題がある．この課題を解決するため，ドローンに搭載したマイクロホンアレイで収録した音情報を利用した探索手法についての研究が行われてきた [6–11]．このような手法はドローン聴覚と呼ばれ，ドローンに搭載したマイクロホンアレイを用いて収録した人由来の音源に基づいた音源定位による，要救助者の探索活動への応用が検討されてきた．このようなドローン聴覚の実装には，マイクロホンアレイや小型処理ユニットといった必須機器をドローンに搭載するための治具が必要となる．このような治具はドローンのペイロードを満たすよう，特定のドローン本体に直接固定するような最小限の設計となっており，各ドローン専用に開発されてきた．したがって治具を個々のドローンの形状や寸法に合わせて再設計する必要があり，汎用性が低い．そのためドローン聴覚の導入において，治具の設計コストと製造コストが必要となるという課題がある．

著者らは以前の研究において，多くの産業用ドローンに一般的な逆 T 字型のランディングギアに着目し，様々なサイズのドローンに容易に取り付けることができる治具を提案した [12]．具体的には，提案した治具はドローン本体ではなくランディングギアに取り付けるよう，最小限の調整で異なるサイズのドローンに設置することができる設計とした．この治具について次の 4 つの指標に基づいて設計し，評価した．第一に製作した治具の質量を計測し，複数のドローンのペイロードと比較することで質量を評価した．第二に有限要素法（FEM）に基づいた静荷重解析により耐荷重を評価した．第三に FEM を用いた固有値解析により振動特性を評価した．第四に実際に製作した治具を複数のドローンに取り付けることで調整の容易さを評価した．しかしながら，これらの評価のうち耐荷重と振動特性の評価は FEM による数値解析に限られているという課題がある．耐荷重の評価は，搭載する計測機器の質量に基づいて静荷重を設定し，十分な安全率を考慮した静荷重解析により適切に評価することができると考えら

*連絡先：東京科学大学工学院機械系
〒 152-0033 東京都目黒区大岡山 2-12-1
E-mail: tsukamoto.y.ak@m.titech.ac.jp

れる．一方振動特性は，実際に製作した治具の寸法や組み立て誤差，さらには締結部による影響があると考えられる．したがって，治具単体での固有値解析に加え，実際の飛行条件下での振動特性の評価が必要であると考える．

したがって本研究では，ドローン単体の場合，ドローン本体底部に従来の治具を直接取り付けた場合，ドローンのランディングギアに提案する治具を取り付けた場合の 3 種類の条件について，屋外でのドローンの飛行中と離着陸時の加速度応答を計測した．そして従来の治具と提案する治具の各点における加速度応答を比較し，実際の飛行条件下での治具の振動特性を評価した．

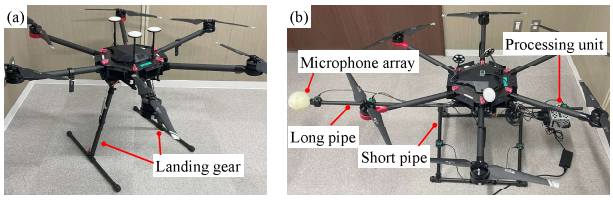


図 1: ドローンと治具の外観： (a) ドローン単体，(b) 提案する治具を取り付けた状態．

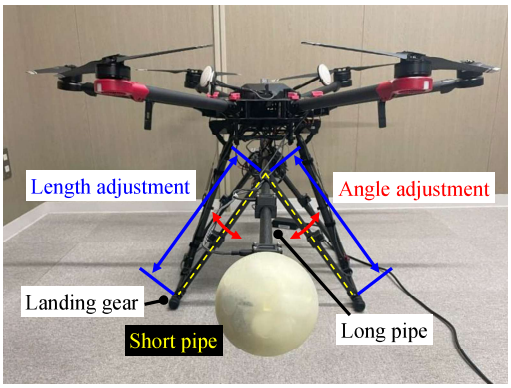


図 2: 提案する治具を取り付けた状態の正面図．

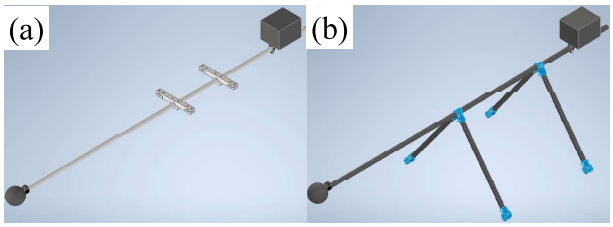


図 3: 固有値解析に用いたモデル：(a) 従来の治具，(b) 提案する治具．

2 治具の設計

多くの産業用ドローンは，図 1(a) に示すような逆 T 字型のランディングギアを有している．そのため図 1(b) に示すようにランディングギアに取り付けることができ，計測機器を搭載することができる治具を設計することで，汎用性の高い治具を実現することができる．そこで図 2 に示すように，トラス状に配置した短いパイプの一端をランディングギアと接続し，もう一端で計測機器を搭載するための長いパイプと接続するような治具を提案した．この設計の特徴は，短いパイプと長いパイプを接続するための部品の取り付け角度を調整し，短いパイプの長さを調整することで，様々なサイズのドローンに適用することができるという調整可能性である．

このような治具を実現するためには，質量，耐荷重，振動特性，調整の容易さの 4 つの設計指標を満たす必要がある．

1. 質量
   治具と計測機器の質量が，ドローンのペイロードを満たす必要がある．

表 1: 数値解析に使用した機械的特性．

	CFRP パイプ	Onyx
密度 [kg/m^3]	1550	1139
縦弾性係数 (E_x) [GPa]	113.6	1.20
縦弾性係数 (E_y, E_z) [GPa]	113.6	0.85
ポアソン比 (ν_{xy}, ν_{yz}, ν_{zx})	0.3	0.35
横弾性係数 (G_{xy}) [GPa]	43.7	0.67
横弾性係数 (G_{yz}) [GPa]	43.7	0.47
横弾性係数 (G_{zx}) [GPa]	43.7	0.39

表 2: 従来の治具と提案する治具の固有振動数．

	Matrice M600 Pro		Matrice 210 RTK V2
	従来手法	提案手法	提案手法
1 次	14.3	17.3	21.9
2 次	15.5	17.5	22.4
3 次	24.3	32.7	70.7
4 次	27.4	37.9	82.9
5 次	109.5	48.9	134.4
6 次	131.7	118.8	199.9
7 次	165.0	215.0	269.3
8 次	182.8	218.0	277.1
9 次	275.0	281.6	301.1
10 次	280.2	309.0	476.7

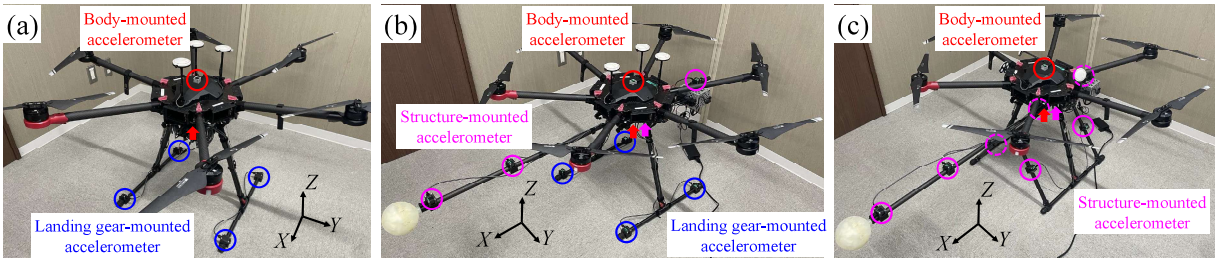


図 4: 加速度センサの配置：(a) ドローン単体，(b) 従来の治具を搭載した場合，(c) 提案する治具を搭載した場合．赤色，青色，マゼンタ色の円は，ドローン本体，ランディングギア，治具に取り付けた加速度センサを示す．

2. 耐荷重
   計測機器の自重に相当する静荷重を治具が支えることができる．

3. 振動特性
   飛行中のロータの振動に由来する共振や疲労破壊を防ぐため，治具が適切な振動特性を有している．

4. 調整の容易さ
   ロータのエゴノイズを低減するため，ドローン本体から離れた位置にマイクロホンアレイを配置する必要がある．さらにドローン全体の重心を調整するため，処理ユニットを長いパイプに沿って調整できる必要がある．そのためこのような計測機器の位置調整を容易に実施できる必要がある．

本研究では，主な構造部品を断面が外径 20 mm，肉厚 1mm の CFRP(Carbon Fiber Reinforced Plastics) パイプとした．さらにこれらのパイプを組み合わせる部品を短繊維の炭素繊維強化ナイロンフィラメント (Onyx, Markforged) [13] を用いた中実の 3D プリント部品とした．

3 数値解析による評価

以前の研究では，前章で示した治具の設計指標の評価は，Matrice 210 RTK V2 (DJI) [14] に取り付けるという条件の下で実施した．本章では，よりサイズの大きいドローンである Matrice M600 Pro (DJI) [15] に取り付けるという条件の下で再度評価した．4 つの設計指標のうち，質量，耐荷重，調整の容易さの 3 つの設計指標は，以前の研究で十分に評価できていると考える．

まず質量は，以前検討した DJI Matrice 210 RTK V2 のペイロードが 1.23 kg であるのに対し，DJI Matrice M600 Pro のペイロードが 5.50 kg であり，治具の質量は大きく変化しないため，質量の設計指標を満たすと考える．次に耐荷重は，同一の計測機器を搭載するため，以前の研究と同様に FEM に基づく静荷重解析により評価できると考える．そして調整の容易さも，以前の研究と同様の構造であるため，適切に評価できていると考える．

一方振動特性は，適用するドローンのサイズに応じて短いパイプの長さと取り付け角度が異なるため，適用対象のドローンに対応した条件で評価する必要があると考える．そのため，適用対象のドローンに対応した形状と寸法の治具について評価する必要があると考える．そこで設計した治具について，FEM 解析ソフト (Inventor Nastran, Autodesk) を用いた固有値解析により振動特性を評価した．

3.1 固有値解析の条件

図 3 に示すように，ランディングギアに取り付けるような提案する治具に加え，比較対象としてドローン本体の底面に直接取り付けるような従来の治具について，マイクロホンアレイと処理ユニットを含む 2 種類の有限要素モデルを作成した．これらのモデルで CFRP パイプはシェル要素，3D プリント部品はソリッド要素と仮定した．従来の治具はドローン本体底面の部品との接触箇所，提案する治具はランディングギアとの接触箇所での固定を拘束条件とした．表 1 に示すように，以前の研究で実施した固有値解析と同様の機械的特性を使用した [12, 16, 17].

3.2 固有値解析の結果

表 2 は，従来の治具と提案する治具を Matrice M600 Pro に適用した場合，提案する治具を Matrice 210 RTK V2 に適用した場合の 3 種類の条件の，1 次から 10 次までの固有振動数を示す．これらの結果から，Matrice 210 RTK V2 に適用した場合の固有振動数は，Matrice M600 Pro に適用した場合の固有振動数と大きく異なり，適用対象のドローンに対応したモデルを用いて評価する必要があることがわかる．

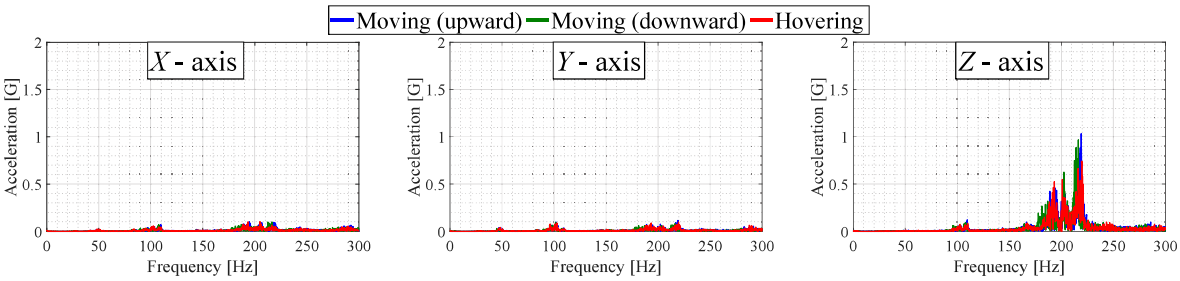


図 5: 上昇，下降，ホバリング時のドローン単体で計測した，ドローン本体上面の加速度応答の周波数スペクトル．

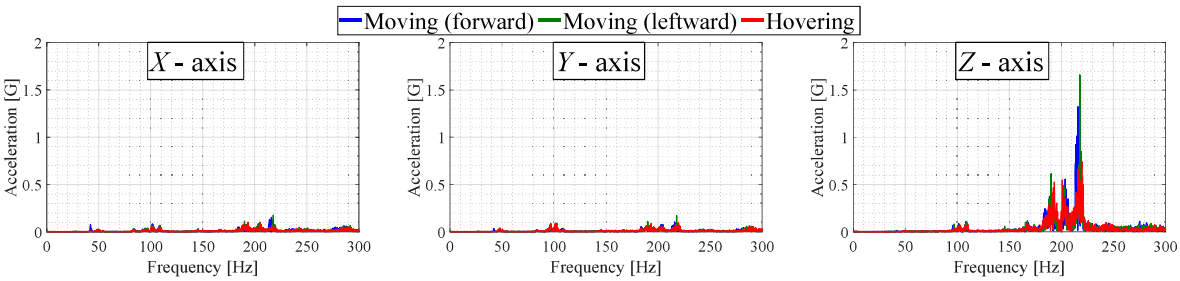


図 6: 前方への移動，左方への移動，ホバリング時のドローン単体で計測した，ドローン本体上面の加速度応答の周波数スペクトル．

4　実飛行条件下での振動特性の評価

本章では，屋外でのドローンの飛行中の治具の振動特性を評価する．この評価では，計測機器を搭載した治具をドローンに取り付け，ドローン本体，ランディングギア，治具上の数か所での加速度応答を計測し，比較することで評価した．

4.1　計測条件

図 4 に示すように，ドローン単体の場合，従来の治具をドローン本体の底面に取り付けた場合，提案する治具をランディングギアに取り付けた場合の 3 種類の条件について振動特性を評価した．各条件について，加速度センサをドローン本体，ランディングギア，治具に配置し，ドローンの対称性を考慮して，上昇，下降，前方への移動，左方への移動，ホバリング，離陸，着陸の 7 種類の飛行条件について加速度応答を計測した．ただし座標軸は，ドローンの進行方向を X 軸，左方を Y 軸，鉛直方向を Z 軸とした．加速度応答の計測条件は，M5Stack ATOM Matrix に搭載されている MPU 6886 を加速度センサとして使用し，サンプリング周波数を 1 kHz，計測時間を 5 秒間として計測したため，周波数分解能は 0.2 Hz として解析した．

4.2　計測結果

得られた加速度応答のうち特徴的な加速度応答について，6 つの観点から比較し，実際の飛行条件下での治具の振動特性を評価する．

4.2.1　ドローンの飛行条件による振動への影響

今回の振動特性の評価で検討するべき飛行条件を明らかにするため，ドローン単体で 5 種類の飛行条件（上昇，下降，前方への移動，左方への移動，ホバリング）で飛行した際に計測した，ドローン本体上面での加速度応答を比較した．図 5 では，青線，緑線，赤線が上昇，下降，ホバリング時，図 6 では，青線，緑線，赤線が前方への移動，左方への移動，ホバリング時の加速度応答の周波数スペクトルを示す．X 軸と Y 軸方向では，どの飛行条件でも振動特性の差異が小さいことがわかる．Z 軸方向でも類似した振動特性を示すものの，220 Hz 付近では飛行条件に応じて加速度応答が異なることがわかる．しかしながら，いずれの方向でも同一のピーク周波数を示しており，飛行条件による振動特性への影響は小さいと考えられる．したがって以降の比較では，ホバリング時の振動特性を代表例として取り上げ，評価を実施する．

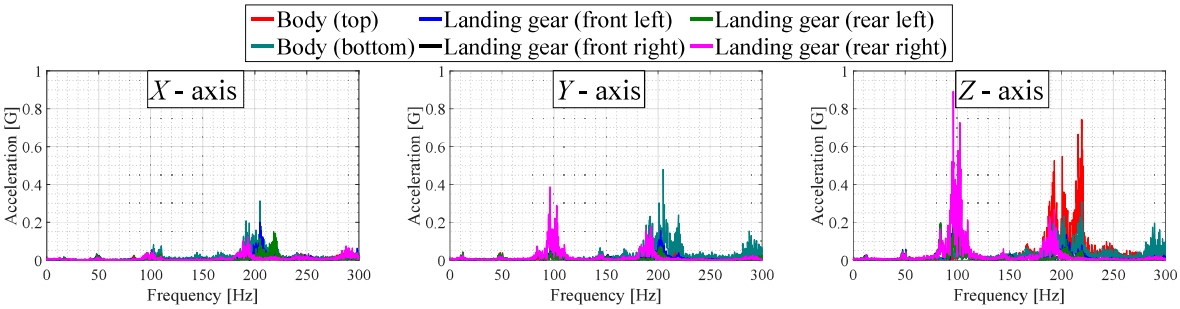


図 7: ドローン単体で計測したドローン本体上面及び底面，ランディングギアの端部における加速度応答の周波数スペクトル．

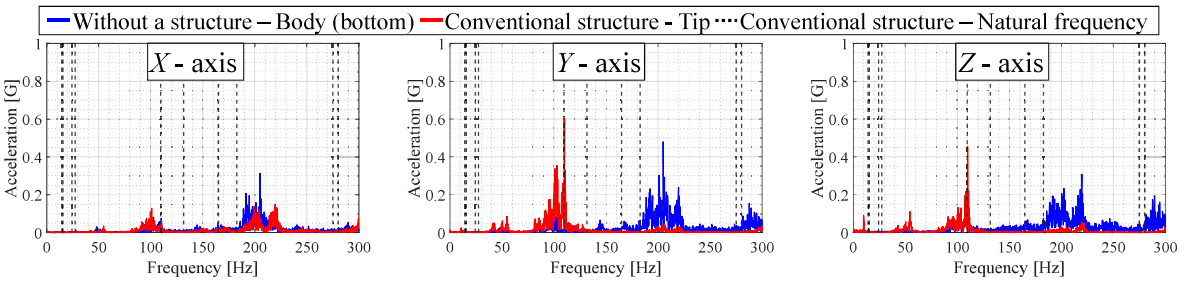


図 8: ドローン単体で計測したドローン本体底面及び従来の治具を取り付けて計測した治具の端部における加速度応答の周波数スペクトル．

4.2.2 ドローン本体の振動とランディングギアの端部における振動の比較

ドローン単体のホバリング時に得られたドローン本体上面及び底面，ランディングギアの端部における加速度応答を比較する．図 7 に，赤線と明るい青線がドローン本体の上面及び底面，青線，緑線，黒線，マゼンタ線がランディングギアの端部における加速度応答の周波数スペクトルを示す．いずれの方向でも，ドローン本体に比べてランディングギアの端部では大きな振動が生じ，Y 軸と Z 軸方向では 100 Hz 付近で大きな振動が生じることがわかる．これはドローンのランディングギアが自由端となり，ロータの振動が増幅されたことが原因であると考えられる．ただし，Z 軸方向では 220 Hz 付近でドローン本体の上面と底面で大きな振動が生じており，ロータの振動による影響であると考えられる．

4.2.3 ドローン本体の振動と従来の治具の端部における振動の関係

ドローン単体で計測したドローン本体底面での振動と，従来の治具をドローンに取り付けて計測した治具の端部での振動との関係を評価した．図 8 に，青線がドローン単体でのドローン本体底面，赤線が従来の治具をドローンに取り付けた際の治具の端部における加速度応答の周波数スペクトルを示す．さらに数値解析により得られた，従来の治具の固有振動数を黒色の点線で示す．いずれの方向でも治具の端部では，ドローン本体の振動を抑制することができている一方で，Y 軸と Z 軸方向では，110 Hz 付近の大きな振動が生じることが確認できる．

この 110 Hz 付近の振動は，治具の固有振動数と一致しており，共振により生じたと考えることができる．さらに今回の比較では，ドローン本体底面で計測した加速度応答を用いており，従来の治具を取り付けた位置で生じる振動とは異なる可能性がある．したがって，治具の振動特性による影響と加速度応答の計測位置の違いにより，ドローン単体で計測したドローン本体底面における加速度応答と従来の治具を取り付けて計測した治具の端部における加速度応答に大きな差異が生じたと考える．

4.2.4 ランディングギアの端部と提案する治具の端部における振動の関係

ドローン単体で計測したランディングギアの端部における振動と提案する治具をドローンに取り付けて計測した治具の端部における振動との関係を評価した．図 9 に，青線，緑線，黒線，マゼンタ線がドローン単体で計測したランディングギアの端部，赤線が提案する

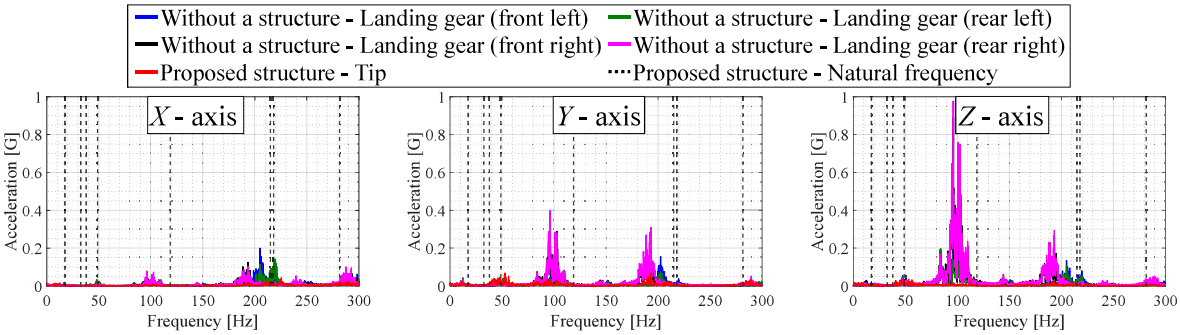


図 9: ドローン単体で計測したランディングギアの端部及び提案する治具を取り付けて計測した治具の端部における加速度応答の周波数スペクトル.

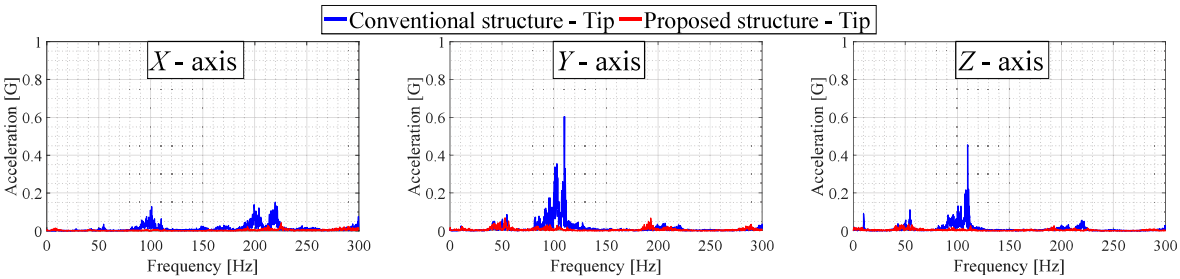


図 10: 従来の治具と提案する治具を取り付けて計測した治具の端部における加速度応答の周波数スペクトル.

治具を取り付けて計測した治具の端部における加速度応答の周波数スペクトルを示す．さらに数値解析により得られた，提案する治具単体での固有振動数を黒色の点線で示す．いずれの方向でも治具の端部では，振動をよく抑制できることがわかる．これはドローン単体でのランディングギアの端部で生じる振動の周波数と提案する治具単体での固有振動数が一致しないことが主な理由であると考える．さらに提案する治具をドローンに搭載することで，ランディングギアと治具がトラス状の構造を構成するため，振動を効果的に抑制できていると考える．

4.2.5 従来の治具の端部と提案する治具の端部における振動の比較

従来の治具と提案する治具をドローンに取り付けて計測した際の，ドローンの飛行中の治具の端部における振動特性を評価する．図 10 に，青線が従来の治具，赤線が提案する治具をドローンに取り付けて計測した，治具の端部における加速度応答の周波数スペクトルを示す．いずれの方向でも，提案する治具は治具の端部における振動を抑制することができている．このような結果は，次の 3 つの要因により生じたと考えられる．第一にランディングギアで生じる振動が，提案する治具単体での固有振動数と一致しないこと．第二に提案する治具が，振動をより効果的に抑制することができる構造となること．そして第三に振動源であるロータからランディングギアまでの距離が遠く，振動の伝播過程で振動が減衰したと考えられる．

4.2.6 ドローンの離着陸による治具の端部の振動への影響

従来の治具と提案する治具をドローンに取り付けて計測した，ドローンの離着陸時の治具の端部における加速度応答を比較，評価した．図 11，図 12 に離陸時と着陸時の結果を示しており，いずれの図においても，青線が従来の治具，赤線が提案する治具を取り付けた場合の加速度応答の結果を示す．離陸時は提案する治具では，従来の治具に比べて一定で小さな加速度応答が得られた．これは離陸時のロータの回転が上昇時の回転と類似しており，ドローンの飛行中と同様にロータの振動が減衰して伝播したことが原因であると考えられる．一方着陸時は，提案する治具では，治具の端部に比べ大きな振動が生じることが確認された．これは提案する治具がランディングギアに直接取り付けられるため，ドローンの着陸時の衝撃が治具の端部にほぼ直接的に作用することが原因であると考えられる．

これらの結果は提案する治具が従来の治具に比べ，ドローンの飛行中は有効であることを示している．さらに図 11 及び図 12 に示すように，提案する治具の端部の飛行中の加速度応答は，いずれの方向でも 1 G 以下

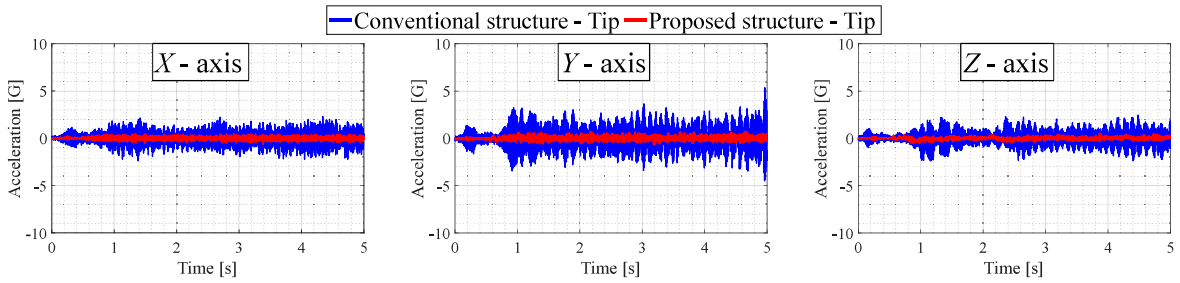


図 11: 従来の治具と提案する治具を取り付けて計測した，ドローンの離陸時の治具の端部における加速度応答.

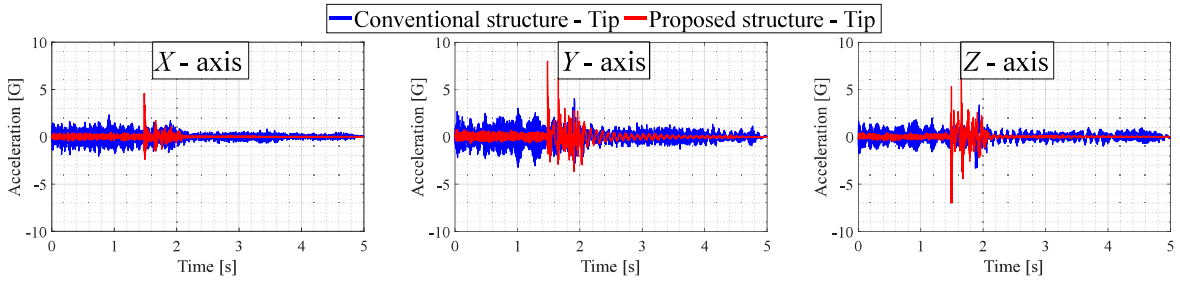


図 12: 従来の治具と提案する治具を取り付けて計測した，ドローンの着陸時の治具の端部における加速度応答.

に抑制することができており，計測機器への影響を最小限にとどめられることが明らかになった．一方で着陸時には，提案する治具の端部では，従来の治具の端部に比べ大きな振動が生じることがわかる．着陸時に音響計測を行わないため，収録音への影響は限定的であるため，構造が十分な強度を有していれば問題ないと考えられる．

5 結言

本研究ではドローン聴覚に用いる計測機器を搭載するための治具のうち，以前の研究で提案した，ドローンのランディングギアに取り付けることができる汎用性の高い治具について，実際の飛行条件下での振動特性を評価した．ドローン単体の場合，従来の治具をドローン本体底面に取り付けた場合，提案する治具をドローンのランディングギアに取り付けた場合の 3 種類の条件について，それぞれ 7 種類の飛行条件で加速度応答を計測した．これらの結果を比較することで，提案する治具の振動特性が従来の治具より優れており，ドローンの飛行中の計測機器への影響を抑制できることが明らかになった．さらに提案する治具は，従来の治具に比べ，着陸時に大きな振動が生じることが明らかになった．このような振動は，着陸時に音響計測を実施しないため，ドローン聴覚による音源定位に与える影響は限定的であるものの，着陸時の衝撃力に耐えることができるよう，十分な強度を有する治具を設計する必要があることが明らかになった．

本研究では，5 m/s 程度の風があるような屋外環境での，実際の飛行中の治具の振動特性を評価した．これらの結果から今後の研究では，同様の屋外環境において飛行中に音を収録し，収録音への飛行中の振動による影響を評価する．さらに同様の評価をより風速の早い環境や騒音が大きい環境で実施し，収録音への影響を評価する．そしてこれらの収録音について，同様の計測地点に配置し，ドローンの飛行による影響がないように地面に対して固定されたマイクロホンアレイで収録した音と比較し，S/N 比といった評価指標を用いて評価する．さらに，今回の結果から着陸時の衝撃への対策が必要であることが明らかになったが，このような着陸時の衝撃を吸収するような機構要素についても検討する必要があると考える．

謝辞

本研究の一部は，JSPS 科研費 22K14218 および NEXCO 関係会社高速道路防災対策等に関する支援基金の助成を受けた．

参考文献

[1] Urbanová, P., Jurda, M., Vojtíšek, T., Krajsa, J.: Using drone-mounted cameras for on-site body documentation: 3D mapping and active survey,

Forensic Science International, vol. 281, pp. 52–62 (2017)

[2] Quan, A., Herrmann, C., Soliman, H.: Project Vulture: A Prototype for Using Drones in Search and Rescue Operations, *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pp. 619–624 (2019)

[3] Sambolek, S., Ivasic-Kos, M.: Automatic Person Detection in Search and Rescue Operations Using Deep CNN Detectors, *IEEE Access*, vol. 9, pp. 37905–37922 (2021)

[4] Mishra, B., Garg, D., Narang, P., Mishra, V.: Drone-surveillance for search and rescue in natural disaster, *Computer Communications*, vol. 156, pp. 1–10 (2020)

[5] Tuan Do, T., Nguyen, D. D., Ho, X. H., Nguyen, V. D.: Leveraging Deep Learning Technique with Meta-Data-Featured Images Captured by Drone for Real-Time Victim Positioning in Flood-Affected Areas, *2025 19th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pp. 1–8 (2025)

[6] Yamamoto, Y., Tang, Z. C., Taguchi, K., Tsuchiya, H.: Drone with Multiple Sensors for Night-time Search Missions of Disaster Victims, *2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA)*, pp. 1-6 (2024)

[7] Wang, L., Cavallaro, A.: Deep-Learning-Assisted Sound Source Localization From a Flying Drone, *IEEE Sensors Journal*, vol. 22, no. 21, pp. 20828–20838 (2022)

[8] Clayton, M., Wang, L., McPherson, A., Cavallaro, A.: An Embedded Multichannel Sound Acquisition System for Drone Audition, *IEEE Sensors Journal*, vol. 23, no. 12, pp. 13377–13386 (2023)

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

[10] Hoshiba, K., Nakadai, K., Kumon, M., Okuno, H. G.: Assessment of MUSIC-Based Noise-Robust Sound Source Localization with Active Frequency Range Filtering, *Journal of Robotics and Mechatronics*, vol. 30, no. 3, pp. 426–435 (2018)

[11] Hoshiba, K., Komatsuzaki, I., Iwatsuki, N.: Proposal of Practical Sound Source Localization Method Using Histogram and Frequency Information of Spatial Spectrum for Drone Audition, *Drones*, vol. 8, no. 4, 159 (2024)

[12] Tsukamoto, Y., Hoshiba, K., Iwatsuki, N.: Development of a Versatile Structure for Mounting Drone Audition-purposed Measurement Equipment, *2025 IEEE/SICE International Symposium on System Intergation (SII)*, pp. 1028–1033 (2025)

[13] Markforged, "Onyx" , `https://markforged.com/jp/materials/plastics/onyx` (accessed on November 11, 2025)

[14] DJI, "Support for Matrice 200 Series V2", `https://www.dji.com/support/product/matrice-200-series-v2` (accessed on November 11, 2025)

[15] DJI, "Support for Matrice 600 Pro", `https://www.dji.com/support/product/matrice600-pro` (accessed on November 11, 2025)

[16] HOPEC, "Technology", `https://www.hopec.jp/technology/` (accessed on November 11, 2025)

[17] TORAY, "Mechanical properties", `https://www.toplaseiko.com/oshidashi/data/` (accessed on November 11, 2025)

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-068-07

ドローン自己雑音のビームフォーミング解析に基づく地表材質推定の検討

Groud Surface Material Estimation Based on Beamforming Analysis of Drone Self-Noise

矢野 翼 [1*]　Benjamin Yen[1,2]
芦澤 剛 [1]　中臺 一博 [1]

[1] 東京科学大学
[1] Institute of Science Tokyo
[2] 理化学研究所
[2] RIKEN

Abstract: 本稿では，ドローン自身が発するローターノイズを活用して，ドローン直下の地表材質を推定する手法を扱う．従来のデータセットは，環境の多様性や収録条件に制約があり有効なデータが限られていたため，本研究では砂利や雪を含む 6 種類の地表材質に対して約 3 時間分の音響データを新たに収録し，データセットの拡張を行った．さらに，地表方向へのビームフォーミングの適用可能性を検討し，深層学習モデルによる材質分類を行った．本研究は，ローターノイズを抑圧すべき雑音としてではなく有用な信号として積極的に活用するという，新たなドローン聴覚のアプローチを提示し，災害現場など屋外環境における堅牢なセンシング技術への貢献を目指すものである．

1　はじめに

近年，豪雨や地震などの自然災害が頻発し，土砂崩れや河川の氾濫によって，道路の寸断や孤立集落の発生といった深刻な被害が生じている．2019 年の台風 19 号では，東日本を中心に 20 都県で 950 件を超える土砂災害が発生し [1]，2020 年の熊本豪雨では，国道の複数個所で通行止めとなった．2024 年の能登半島地震では，県管理の道路だけで 90 か所以上が通行不能となり，孤立集落の把握にも時間を要した [2]．こうした災害時には，道路や地表の状態を迅速かつ広域に把握することが復旧支援に不可欠である．特に，道路が泥に覆われているか，アスファルトが露出しているかといった地表の材質の違いは，通行可否や被害の深刻度を判断する上で重要な情報となる．しかし，従来の視覚ベースの画像認識手法では，泥濘や影，照度変化の影響により判別が困難な場合がある．そこで本稿では，音響情報を用いた地表材質推定に着目する．特に，ドローンに搭載したマイクロホンで収録する音響信号には，ローターからの直接音に加え，地表で反射した反射音も含まれており，この反射音には地表材質に応じた特徴が現れる．本稿では，この反射音を活用した地表材質推定方法を提案するとともに，屋外環境における推定性能を評価するためのデータセットの構築も行う．

*連絡先： 東京科学大学
152-8552 東京都目黒区大岡山 2-12-1
E-mail: yano@ra.sc.eng.isct.ac.jp

2　関連研究

災害現場における地表材質の把握は，被害状況の評価や復旧計画の策定において重要であるが，視認性やアクセス性の制約から，従来の視覚ベースの方法では限界がある．そこで，近年ではドローンにセンサを搭載し，被災状況を把握する研究が進められており，特にカメラやマイクロホンを用いた手法が提案されている [3]．カメラは，有効な観測手段だが，夜間や煙，霧など視界不良下では利用が困難という課題がある．これに対し，マイクロホンを用いた音響センシングは，視界に依存せず環境情報を取得できる利点から,「ドローン聴覚」として注目され，多くの研究が行われている [4, 5, 6, 7, 8, 9]．ただし，既存研究の多くは，被災者の音声検出や位置推定を目的としており，ローターノイズは抑圧すべき対象とされてきた．一方，本稿では，従来ノイズとみなされていたローターノイズに含まれる地表からの反射音に着目し，それを積極的に活用する点で，従来研究と一線を画す．図 1 に示すように，反射特性の違いから地表材質の変化を検出できれば，災害による地表面のの変化（例：土砂の堆積，道路の損壊など）を把

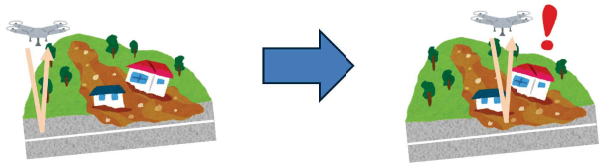


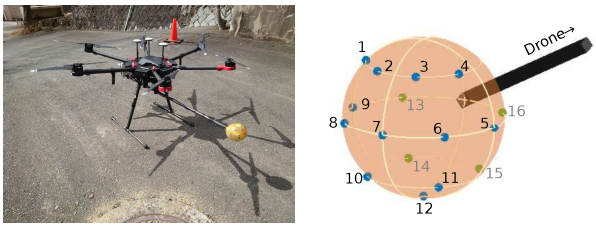
図 1: 本稿の応用例のイメージ：反射音から地表面材質を推定する．既存地図と材質が異なる部分があれば，被災により地表面に何らかの影響あったことが分かる．

握できる．このようなアプローチは，われわれの知る限りでは未踏の研究領域である．

本稿では，反射音が地表面方向から到来する点に着目し，信号処理分野におけるビームフォーミング技術 [10] を適用することで反射音特徴のみを抽出し，性能向上を図る．ビームフォーミングは，複数のマイクロホンから得られる信号に時間遅延や重みづけを施すことで，特定方向に対するビームを形成し，その方向から到来する音を強調する技術であり，音源定位，音源分離，ノイズ抑圧など様々な応用が実用化されている [11]．地表面からの反射音が特定方向から到来する状況では，その方向にビームを形成することで，反射成分を選択的に抽出・強調することが可能となる．

また，音響信号を用いた材質推定の事例としては，橋梁の打音検査があげられる．従来は，熟練作業者が橋梁を打撃して音を聞き分けていたが，近年では，これを機械学習で代替する研究が進んでおり [12]，一部では，打音検査機能を有する付着走行型ドローンも報告されている [13]．ただし，これらは，主にコンクリート構造物を対象としており，多様な地表材質を対象に，反射音のみから材質を識別する研究はほとんどない．また，打音検査では，人工的に音を発生させる必要があるが，本稿では，ドローンローターから自然に発生するノイズを情報源として活用しており，従来手法とは根本的に異なるアプローチをとっている．このように，本研究は従来「抑圧すべき雑音」とされていたローターノイズを「有用な情報源」と再定義し，災害対応や野外センシングにおける音響情報の新たな活用の可能性を拓くものである．

われわれは以前，屋外環境でドローン飛行中の音響データを収録し，深層学習により地表材質を分類可能であることを示した [14]．しかし，当時のデータセットは強風の影響で有効なデータが限られ，クラス数・量ともに不十分だった．さらに，各材質につき 1 回のフライトのみの収録であったため，反射音よりも，各フライトに特有のノイズ成分を学習してしまう可能性もあった．そこで本稿では，データセットの拡張を行うとともに，データの処理が行いやすいような，外部ノイズの少ないデータ取得を狙う．また，各素材に対して複数回のフライトを行い，汎化性能の向上を図るほか，前述のビームフォーミング技術を地表材質推定手法に取り込むことにより，外部環境に対する頑健性の向上を狙う．



(a) 機器取り付け後のドローン　(b) アレイチャンネル配置

図 3: 収録に用いるドローンとマイクアレイ

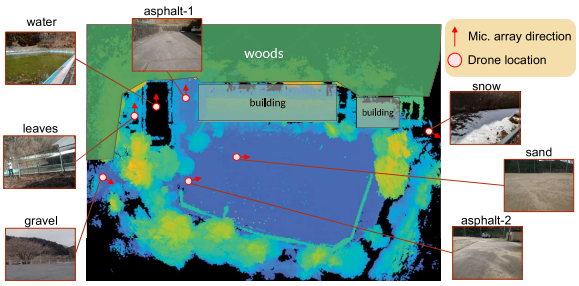


図 4: 各収録場所と地表面の様子

3　データセットの拡張

本稿では，地表材質推定に用いる音響データの信頼性と汎化性能を高めるため，新たな屋外データセットの収録を行った．本章では，その収録方針，手法，得られたデータの分析結果について述べる．

3.1　新規データ収録の指針

マイクロホンで録音されるローターノイズには，地表の材質だけでなく，飛行高度，各モーターの回転速度，地形など，様々な要素が影響を与える．本稿では，既存のデータセットが抱えていた課題を踏まえ，以下の方針に基づいて新たなデータを収録した．

1. **パラメータの固定化**：地表材質とドローン高度以外の要素（例：ドローンの機体設定や環境条件）を可能な限り統一する．
2. **材質の選定**：災害現場を想定し，現実的に出現しうる地表材質を選択する．
3. **高度の設定**：高度が高すぎると反射音が減衰し，推定が困難になる．風の影響も大きくなるため，人の頭上よりやや高い程度（3〜5 m）に限定．
4. **データ量の確保**：バッテリの制約内で，各材質に対してできる限り長時間の収録を行う．
5. **複数回収録による汎化性向上**：1 つの地表材質につき，複数回のフライトを行う．

表 1: 新規収録データ内訳

収録場所	収録時間長（フライト回数）
砂	79 分（6）
アスファルト 1	19 分（2）
アスファルト 2	20 分（2）
水	20 分（2）
砂利	20 分（2）
落ち葉	20 分（1）
雪	9 分（1）

表 2: データセット作成条件

使用機体	DJI Matrice 600
サンプリング周波数	16 kHz

3.2 データ収録手法

収録に用いた機体について，先行研究 [14] と同様，図 3 に示すように，16 ch の球形マイクロホンアレイをドローンの前方向に突き出す形で設置し，後方には，録音装置として Raspberry Pi を搭載した．機体の重心を考慮して，それぞれの位置は慎重に調整・固定した．この機体をホバリング状態にして，地表の材質を都度変えながら，取り付けたマイクロホンアレイの全てのチャンネルで録音を行った．収録対象の地表材質は．砂，アスファルト，水，砂利，落ち葉，雪の 6 種類とし，アスファルトについては周辺環境と広さが異なる 2 地点（アスファルト 1 と 2）を用意して，環境変化に対する汎化性能を評価可能にした．これらの地表材質を有する地点はいずれも大きな凹凸はなく，傾斜に関しても雪以外はほぼ水平であった．各収録地点の配置と，地表面の様子を図 4 に示す．図中の色は標高の相対差を示しており，青は低地，黄色は高地（木立や人工物のある場所）を表す．また，図中の小円はドローンの位置であり，赤い矢印の方向がマイクアレイの向きである．これらは風による影響を抑えるため，いずれも風下側を向くように設定した．収録は，各材質の上空でホバリングしながら，10 分間，もしくは 20 分間実施した．ただし，気象やドローンのバッテリ状況によって，各回の収録時間には数分程度のばらつきがある．収録時の風速は，平均して 3 m/s 程度であった．最終的に約 3 h 分のデータを収録することに成功した．その詳細を表 1 に，その他の条件を表 2 にまとめている．

3.3 録音データの分析

収録された音響データの品質と一貫性を評価するため，既存データと新規データの比較分析を行った．具体的には，録音中に生じる不具合の有無や，チャンネル間での音響特性のばらつきを調査し，今後の推定精度向上に向けた課題と改善点を明らかにする．

3.3.1 不具合発生の確認

既存の屋外環境データには,「音割れ」と,「録音値が 0 になる」といった不具合が含まれていた．本稿では，新規に収録したデータと既存データを比較し，こうした問題の発生頻度を調べ，表 3 にまとめた．表は，各収録場所での録音データについて，全フライトでチャンネルごとに単位時間での不具合観測回数を求め，その最大値を代表値として記載している．この結果，音割れは一部の地点で従来の 10 倍以上の頻度で発生していた一方で，録音値が 0 となる区間は大幅に減少していた．音割れの多くは，風が吹き抜けやすい場所（例：砂，アスファルト 2）で発生しており，上空では地上よりも強い風が吹いていたこと予想される．逆に木立や建物に囲まれた場所（アスファルト 1，落ち葉）では比較的発生頻度が低かったため，強い風が吹き抜けにくかったと考えられる．また，録音値 0 の区間が減少した理由としては，マイクアレイのオーバーホールにより，録音値が 0 となる現象の主要因であった振動板が内壁に貼り付く頻度が少なくなったためと考えられる．

3.3.2 チャンネル間の比較

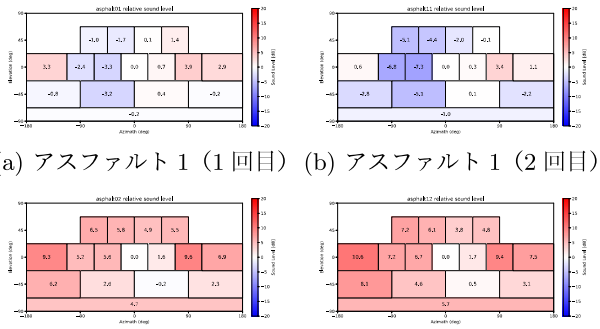
既存のデータセットでは，各素材につき 1 回のフライトのみでデータを収録していたため，ネットワークの学習時に，複数のチャンネルのデータをまとめて入力すると，学習したい地表面からの反射音よりも，そのフライト固有の環境要因が学習されてしまう懸念があった．そこで本稿では，新たに収録したデータに対して，チャンネル間の音圧レベルを比較し，同一材質に対する録音の一貫性を検証した．図 5 は，アスファルトの上空で収録した 4 つのデータにおける，最前方の第 7 チャンネルを基準とした音圧差を示している．理想的には，第 7 チャンネルがドローンローターから最も遠いため，音圧も最も低くなると考えられるが，アスファルト 1 では機体から見て右前方の音圧レベルが低くなるなど，アスファルト 2 とは異なる傾向がみられた．これは風向きや周囲の建物など，環境の違いによる影響と考えられる．これは，すべてのチャンネルのデータをまとめて入力すると，ネットワークがフライト固有の環境要因を学習してしまうという前述の仮説を支持する結果である．したがって，今後は反射方向に着目したビームフォーミングなどの音処理技術の活用が必要であるといえよう．

4 地表材質推定手法

前節の分析結果から，チャンネルごとの音圧レベルには風向や周辺環境が大きく影響し，単純なチャンネ

表 3: 不具合発生頻度比較

区分	収録場所	音割れ	データ値 0
新規	砂	214.2	0.8
	アスファルト 1	12.3	0.0
	アスファルト 2	152.3	5.6
	水	50.9	0.0
	砂利	98.7	2.2
	落ち葉	22.7	0.0
	雪	28.0	0.0
既存	土砂	0.2	32.9
	アスファルト	0.2	15.8
	浅い水	4.1	24.0
	深い水	20.5	13.1



(a) アスファルト 1（1 回目） (b) アスファルト 1（2 回目）

(c) アスファルト 2（1 回目） (d) アスファルト 2（2 回目）

図 5: 相対音量比較

ル統合では反射音の特徴を正しく学習できない可能性が示唆された．そこで，本節では，地表からの反射音を抽出するために，録音データに対してビームフォーミングを適用する．使用する手法には，遅延和ビームフォーミング法を選択した．この手法は，各チャンネルの信号に対し，目的方向からの到来音に合わせて時間遅延を補正し，それらを加算することで，その方向からの音を強調することができる．総チャンネル数を M とし，第 m チャンネルの録音信号を x_m，対応する時間遅延を τ_m とすると，遅延和ビームフォーミングによって抽出される目的信号 $x_{DS}(t)$ は以下で表される：

$$x_{DS}(t) = \sum_{m=1}^{M} x_m(t + \tau_m)$$

本稿では，方位角 180°，仰角 -75° の方向に固定してビームを形成し，その方向に対応する時間遅延パラメータを用いた．

材質推定モデルは，先行研究 [15] を踏襲し，4 層構造の CNN（図 6）と，ResNet18 [16] を用いた．入力データ作成にあたり，ビームフォーミング後の録音信号に対してフレーム長 512, シフト長 128 でハニング窓を用いて短時間フーリエ変換 (STFT) を行い，複素スペクトログラムに変換し，直流成分を取り除く．その後，長さ 256 フレーム，シフト長 128 フレームでスペクトログラムをスライスすることで，入力データを得る．スライスした各スペクトログラムの実部と虚部

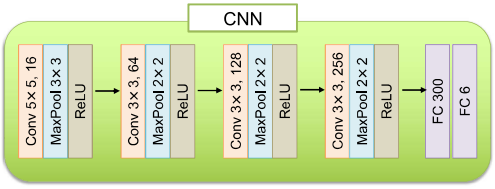
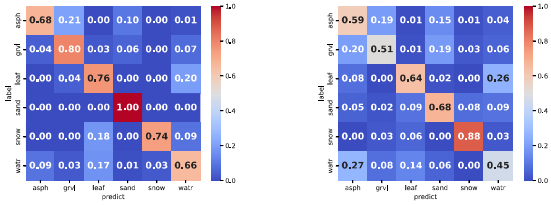


図 6: 本実験で用いる CNN アーキテクチャ



(a) CNN の場合 (b) ResNet18 の場合

図 7: 混同行列のヒートマップ

の 2 チャンネルを入力として利用する．入力サイズは，2 × 256 × 256 となる．

5 評価実験

構築した地表材質推定手法の有効性を確認するため，評価実験を行った．まず，各録音データに対して，遅延和ビームフォーマーを適用し，もとの録音データ中に不具合が確認されていた区間を除去した．その後，処理済み録音データを学習用，検証用，テスト用の 3 つの区間に分割し．全ての区間に対して 4 節で述べた手法により，ネットワークへの入力用の特徴量を作成した．この処理により，学習データ 4097 個，検証データ 525 個，テストデータ 536 個が得られた．学習データには，地表材質に対応する 6 次元の one-hot ベクトルを教師ラベルとして付与した．なお，表 4 に CNN と ResNet18 の学習条件を示す．学習曲線を確認し，過学習が起きていない範囲で，最も性能が良好なエポックのモデルを評価に用いた．性能指標には，以下に示す正答率 (Accuracy) を用いた．

$$\text{正答率 (\%)} = \frac{\text{推定ラベルと正解ラベルが一致したデータ数}}{\text{入力データ数}} \times 100$$

表 5 に各ネットワークにおける学習・検証・テストデータに対する正答率を示す．また，混同行列のヒートマップを図 7 に示す．CNN では，86 エポック目のモデルのテスト正答率が 70.8% となり，ResNet18 では，276 エポック目のモデルを用いて，62.3% の正答率を達成した．

実験結果の詳細を踏まえて考察を行う．CNN のテストデータに対する正答率は 70% を越えているものの，過学習が起きていない範囲では学習が収束しきれておらず，汎化性能の観点で少々疑問が残る．一方，ResNet18 はテストデータに対し 60% 程度と，CNN と比べて劣

表 4: 学習条件

損失関数	クロスエントロピー
オプティマイザ	Adam
学習率	10^{-4}
エポック数	300

表 5: 各データに対する正答率

ネットワーク	学習データ	検証データ	テストデータ
CNN	87.8%	72.2%	76.7%
ResNet18	100%	62.3%	60.2%

るものの，6 クラス分類のチャンスレートと比べて十分高い値であることから，ある程度反射音の特徴を学習できていることが考えられる．今後はデータ区間の割り当ての変更や，既存データと混合しての検証などを通じて，汎化性能について詳しく分析を進める予定である．図 7 に示した混同行列のヒートマップを確認すると，いずれのネットワークも水に対する識別性能がほかの素材に比べて低い傾向が見られた．この原因として，プロペラによる風が水面に影響を与え，反射音の特性が不規則に変動することがあげられるが，現時点では仮説に過ぎず今後の追加分析が必要である．また，収録時に強風の影響で使用できなかったデータが存在したため，実用化に向けては，マイクロホンアレイへの風防の取り付けなど，風への対策が不可欠である．

6 おわりに

本稿では，ドローンのローターノイズに含まれる反射音を利用し，頑健な地表材質推定手法の構築を目指した．屋外環境下で，ローターノイズと地表材質 6 種類（砂，アスファルト，水，砂利，落ち葉，雪）に関するデータセットを作成し，これを対象に CNN および ResNet18 を用いた材質推定モデルを構築した．また，収録したデータに遅延和ビームフォーミングを適用したのち，地表材質推定手法の評価を行い，ResNet18 で正答率 60.2% を達成し，その有効性を示すことができた．外部ノイズや環境に対する具体的な性能の検証や，識別性能の分析，さらなるデータセットの拡張，リアルタイム処理などが，今後の課題である．

参考文献

[1] 国土交通省　砂防部：“令和元年台風第 19 号に伴う土砂災害の概要　ver2.1”，https://www.mlit.go.jp/river/sabo//jirei/r1dosha/r1typhoon19_gaiyou191224r.pdf

[2] 石川県：“令和 6 年（2024 年）能登半島地震に関する情報 被害等の情報について（第 12 報）”，https://www.pref.ishikawa.lg.jp/saigai/documents/202401041500higaihou.pdf

[3] 松井他：“ドローン搭載カメラからの RGBD 画像に対するセグメンテーション等による土砂災害箇所と規模の推定法の検討”, 研究報告コンピュータビジョンとイメージメディア (CVIM), no. 16, pp. 1-6, 2023.

[4] K. Okutani et al.：“Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter”, IEEE/RSJ IROS, pp. 3288-3293, 2012

[5] K. Hoshiba et al.：“Design of UAV-Embedded Microphone Array System for Sound Source Localization in Outdoor Environments”, Sensors, 17(11), 2017

[6] T. Yamada et al.：“Placement Planning for Sound Source Tracking in Active Drone Audition”, Drones, 7(7), 2023

[7] M. Kumon et al.：“Alternating Drive-and-Glide Flight Navigation of a Kiteplane for Sound Source Position Estimation”, IEEE/RSJ IROS, pp. 2114-2120, 2021

[8] B. Yen et al.：“A Performance Assessment on Rotor Noise-Informed Active Multidrone Sound Source Tracking Methods”, Drones, 8(6), 2024

[9] L. Wang et al.：“Drone audition for bioacoustic monitoring”, Methods in Ecology and Evolution, 14(12), pp. 3068-3082, 2023

[10] Van Veen, B. D. and Buckley, K. M.：“Beamforming: a versatile approach to spatial filtering”, IEEE ASSP Magazine, 5(2), pp. 4-24, 1988

[11] P.Chiariotti et al.：“Acoustic beamforming for noise source localization – Reviews, methodology and applications”, Mechanical Systems and Signal Processing, vol. 120, pp. 422-448, 2019

[12] 江本他：“AI 手法による打音検査の浮き判定の検討”, AI・データサイエンス論文集, 1(J1), pp. 514-521, 2020

[13] 上半：“ドローンを用いたコンクリート橋部材の詳細検査手法”, コンクリート工学, 57(9), pp. 699-704, 2019

[14] 矢野他：“屋外環境下でのドローンのローターノイズによる地表材質推定に向けた手法の検討およびマイクロホンアレイ用風防の設計”, SIG-Challenge, 2024

[15] 矢野他：“ドローンのローターノイズによる地表材質推定手法の検討”, SICE SI, 2023

[16] K. He et al.：“Deep Residual Learning for Image Recognition”, CVPR, pp. 770-778, 2016

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-068-08

幾何音響シミュレーションと機械学習を用いた2次元領域に対する波動伝搬予測手法の検討

Study on Wave Propagation Prediction Method for a Two-Dimensional Domain Using Geometrical Acoustics Simulation and Machine Learning

芦澤　剛 [1*]　Yen Benjamin[2,1]　中臺 一博 [1]

[1] 東京科学大学 システム制御系
[2] 理化学研究所 生命機能科学研究センター
[1] Department of Systems and Control Engineering, Institute of Science Tokyo
[2] RIKEN BDR

Abstract: 音響 AI モデルの学習を行う際には，音の波動的な振る舞いが表現されているデータが必要となるが，実環境において多様な条件でのデータを収集することは現実には困難である．また，有限差分時間領域（FDTD）法をはじめとする波動音響シミュレーションを実施する場合でも，大規模な解析モデルに対して高周波領域まで解析しようとすると，依然として計算コストが高い．そこで本研究では，幾何音響（GA）シミュレーションの一つである音線法（Ray Tracing 法）を入力とし，画像間変換モデルを用いて波動音響シミュレーションの結果を予測する Ray2Wave を提案する．本稿では，提案手法の有効性を検証するために，2 次元の矩形モデルに対して，音源や障害物の異なる 20 ケースに対して Ray Tracing と FDTD のペア画像を生成し，そのうち 16 ケースを学習データ，4 ケースをテストデータとして用いた機械学習を行った．今回学習した機械学習モデルでは，音線が存在する領域では全体的な伝搬挙動を再現でき，障害物がない解析モデルに対しては推定誤差が小さかった．その一方，障害物を有する解析モデル対する結果では，特に障害物周囲での推定精度が低下した．これは Ray Tracing の解析結果が回折を表現できていないことに起因すると考えられる．今後はこのようなエラーに着目して，機械学習モデルの改善に取り組む予定である．

1　はじめに

ロボット聴覚における音声認識や音源定位といったタスクは，AI や機械学習の発展によって進歩している．機械学習モデルの学習を行う際には，音波は反射や回折，干渉などの影響を受けるため，これらの波動的な振る舞いを踏まえたデータが必要となる．しかしながら，例えば屋内の限られた条件下での学習データは計測可能である一方 [1]，多様な室環境や，さらに屋外まで想定した多様な環境下での大規模なデータセットを実測にて収集することは，あまり現実的ではない [2, 3]．そこで，シミュレーションで機械学習モデルの学習を行い実環境に適用する simulation-to-real（Sim2Real）が有効な代替手段となり得る [4]．

数値音響シミュレーション手法は，大きく幾何音響シミュレーションと波動音響シミュレーションに分類される．音線法（Ray Tracing Method）や虚像法（Image source method）に代表される幾何音響（GA）解析手法は計算コストが低い一方で，波動現象を考慮しない [5, 6, 7]．これに対して，FDTD，FEM，BEM などの波動音響（WBA）は波の伝搬を精密に再現できるが，計算コストが高い [8, 9, 10, 11]．近年の計算資源の発展により，数キロヘルツ帯までの大規模三次元空間の解析が可能になりつつあるが，依然として GA に比べて高コストである．また，低周波数帯域を WBA，高周波数帯域を GA で扱うハイブリッド手法も提案されているが，高周波帯域に関してはあくまで GA の精度に依存することとなり，また，WBA の実施コストを削除できるものではない [12, 13]．

そこで本研究では，幾何音響の結果を入力として，画像変換モデルにより波動伝搬を予測する手法 Ray2Wave を検討する．

本論文の貢献は以下の通りである．

- 低コストな音波の表現として Ray Traycing によ

*連絡先： 東京科学大学
152-8552 東京都目黒区大岡山 2-12-1
E-mail: ashizawa@ra.sc.eng.isct.ac.jp

る解析結果をモデルの入力とすることで，簡易な機械学習モデルでも波動伝搬挙動を予測することを試みた．

- 2 次元矩形領域に対して異なる音源位置や障害物を定義して解析を行い，U-net を基本とした機械学習モデルを作成した．
- 障害物のないモデルに対しては比較的精度よく予測できたが，障害物のある場合には正しく予測できなかった．これは Ray Tracing が回折を表現できていないことに起因すると考えられ，今後は機械学習モデルの改善が必要である．

2 関連研究

本節では，音響シミュレーションと機械学習を組み合わせた先行研究について確認する．

初期の研究では，深層ネットワークを用いて波動シミュレーションを近似したり [14]，明示的な物理モデルを用いずに振動場を再構成したりする手法が提案された [15]．その後，支配方程式をネットワークに埋め込む枠組みとして PINN（Physics-Informed Neural Networks）が登場し [16]，SIREN に代表される暗黙的ニューラル表現は，高精細な表現能力を持つことが示された [17]．また，ニューラルサロゲートを用いた波動音響シミュレーションの高速化 [18, 19, 20]，PINN による波動方程式の直接解法 [21, 22]，計測データからの音場推定 [23, 24, 25, 26]，さらにはインパルス応答 (IR) 生成やレンダリングに関する研究 [27, 28] も進展している．

中でも本研究と特に関連するものとして以下の 4 つが挙げられる．Ratnarajah らは，GWA データセットを用い，3D メッシュと音源・受音点位置から IR を生成する MESH2IR を提案した [28]．Moseley らは，時空間座標と媒体モデルから長時間の波動場を予測する PINN フレームワークを導入した [21]．Wang らは，複雑構造による散乱場をメッシュ生成なしに PINN で再構成した [22]．Fan らは，物体形状から周波数依存の散乱場を CNN で推定し，計算を大幅に高速化した [29]．

これらの研究は，いずれもサロゲートモデルや特定タスクに対する推定に焦点を置いており，大規模かつ多様な音場に対する一般的な枠組みを対象としてはいない．効率的でスケーラブルな幾何音響は，これまで波動ベース挙動を予測するための中間表現としては利用されてこなかった．Ray2Wave はこの点に着目し，GA を低コストで情報量のある中間表現として活用することで，機械学習モデルの学習が波動特有の現象に集中できるようにすることを試みる．本手法は PINN やサロゲートモデルを置き換えるものではなく，それらと相補的であり，将来的には組み合わせて利用することも可能である．

3 提案手法

3.1 Ray2Wave の概要

図 1 に，本研究で提案する Ray2Wave の概念図を示す．本研究では，入力として幾何音響シミュレーション手法である Ray Tracing 解析を用い，出力として波動音響シミュレーション手法である FDTD を用いた．両者の結果は画像として扱い，image-to-image translation によって変換を行うことを試みた．本節では，解析モデルに加え，Ray Tracing，FDTD，および image-to-image translation の各手順について述べる．

3.2 解析モデル

図 2 に示すように，本研究では 2 次元の矩形領域を解析対象とし，その内部に複数の矩形障害物を配置したモデルを構築した．解析モデルに関するパラメータの詳細を表 1 にまとめる．

音源には，中心周波数 $f_0 = 500\,[\mathrm{Hz}]$ のガウシアンパルスを放射する点音源を使用した．音速は $c_{\mathrm{wave}} = 343\,[\mathrm{m/s}]$ と設定した．空間方向の離散化においては，この中心周波数に基づいて，グリッド間隔 $\Delta x = (c_{\mathrm{wave}}/f_0)\cdot(1/20)\ [\mathrm{m}]$ を用いた．

解析領域は x 方向に 600 グリッド，y 方向に 500 グリッドで構成した．点音源は座標 $(x_{\mathrm{src}}, y_{\mathrm{src}})$ に配置し，障害物は位置 $(x_{\mathrm{obs}}, y_{\mathrm{obs}})$ とサイズ $(w_{\mathrm{obs}}, h_{\mathrm{obs}})$ を持つ矩形として設置した．音源座標・障害物座標・サイズは Table 1 に示す範囲内でランダムに割り当てた．

障害物および解析領域の境界条件は，両解析手法においてサウンドハード境界として設定した．

これらのモデルに対して，幾何音響解析として Ray Tracing 法を，波動音響解析として FDTD 法を適用し，同一条件下で音波伝搬の比較を行った．

3.3 Ray tracing の実施手順

本研究で実施した Ray Tracing による解析は次のとおりである．音源は，点音源の座標から全方位へ 360 本の音線（Ray）を放射することとした．解析は，時間ステップ Δt ごとに音線を逐次的に進めることで行った．障害物および解析領域の境界ではサウンドハード条件を想定し，法線ベクトルに基づいて鏡面反射となるように設定した．各音線については，伝搬の各ステップにおける先端座標を記録し，時系列データとして保存した．

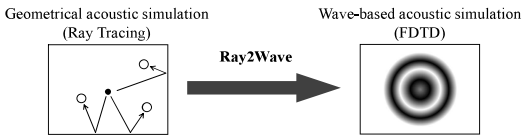


図 1: 提案手法 Ray2Wave の概念図

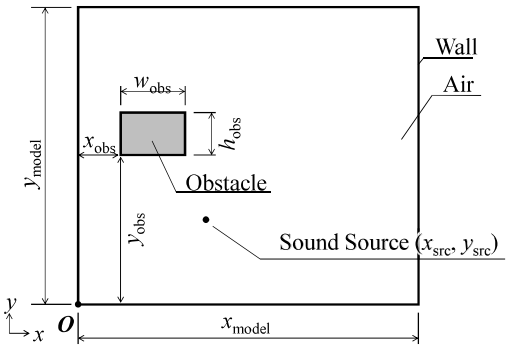


図 2: 解析モデル

表 1: 解析モデルのパラメータ

Center frequency [Hz]	f_0	500
Sound speed [m/s]	c_{wave}	343
Grid spacing [m]	Δx	$(c_{\mathrm{wave}}/f_0)\cdot(1/20)$
Model size [Δx]	$(x_{\mathrm{model}}, y_{\mathrm{model}})$	$(600, 500)$
Source position [Δx]	$(x_{\mathrm{src}}, y_{\mathrm{src}})$	$(50 \sim 450, 50 \sim 550)$
Obstacle position x [Δx]	x_{obs}	$30 \sim (600 - w_{\mathrm{obs}} - 30)$
Obstacle position y [Δx]	y_{obs}	$30 \sim (500 - h_{\mathrm{obs}} - 30)$
Obstacle size [Δx]	$(w_{\mathrm{obs}}, h_{\mathrm{obs}})$	$(30 \sim 100, 30 \sim 100)$
Time Step [s]	Δt	$0.8\,\Delta x/(\sqrt{2}\,c)$

3.4 FDTD の実施手順

次に，本研究で実施した FDTD 法による波動音響解析について述べる．解析は，以下の波動方程式を差分化することで行った．

$$\frac{\partial^2 \phi}{\partial t^2} = c^2 \left(\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} \right) \tag{1}$$

ここで，ϕ は音圧 p および粒子速度 $\mathbf{v}$ と以下の関係を満たすポテンシャルである．

$$p = -\rho_0 \frac{\partial \phi}{\partial t}, \quad \mathbf{v} = \nabla \phi = \left(\frac{\partial \phi}{\partial x}, \frac{\partial \phi}{\partial y} \right). \tag{2}$$

音源には，中心周波数 f_0 のガウシアンパルスを使用した．障害物および解析領域の境界ではサウンドハード条件とするため，Neumann 境界条件を適用した．

3.5 Image-to-Image translation

幾何音響解析（Ray Tracing）によって得られた画像から，波動音響解析（FDTD）による解析結果を予測するために，image-to-image translation モデル [30] を用いることを考える．本モデルは Conditional GAN（条件付き生成的敵対ネットワーク）に基づいて構成され，入力画像（条件画像）から対応する出力画像を生成する．ネットワークの概要を 図 3 に示す．本ネットワークは Generator と Discriminator の 2 つの主要な構成要素からなる．

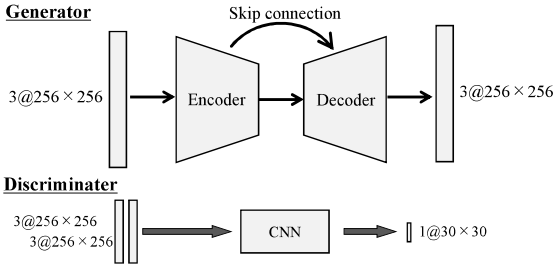


図 3: 画像変換モデル

Generator は 8 層の Encoder と，それに対応する 8 層の Decoder から構成された U-Net である．各層のチャネル数は，入力画像（RGB，3 チャネル）から始まり，Encoder の最深部では 512 チャネルまで増加し，その後 Decoder で順に減少する構成となっている．また，各層はスキップコネクションによって接続されている．

Discriminator には PatchGAN アーキテクチャを採用し，画像をパッチ単位で分類することで，生成画像を評価する．

学習においては，バッチサイズを 4，エポック数を 100 とした．

4 実験

4.1 データセットの作成

本研究では，音源位置に加えて，障害物の数，座標，サイズを変化させることで，合計 20 ケースの解析モデルを作成した．作成した 20 個のモデルを 図 4 に示す．

Case 1 から Case 16 をトレーニングデータ，Case 17 から Case 20 をテストデータとして用いた．各ケースに対して Ray Tracing と FDTD の両方を実行した．

Case 1 における Ray Tracing および FDTD の解析結果の例を図 5 に示す．本実験では Frame #0 〜 #199 の 200 フレームを生成し，出力は $10\Delta t$ ごとに行った．

以上のようにして，トレーニングデータとして $16 \times 200 = 3200$ ペア，テストデータとして $4 \times 200 = 800$ ペアを準備した．

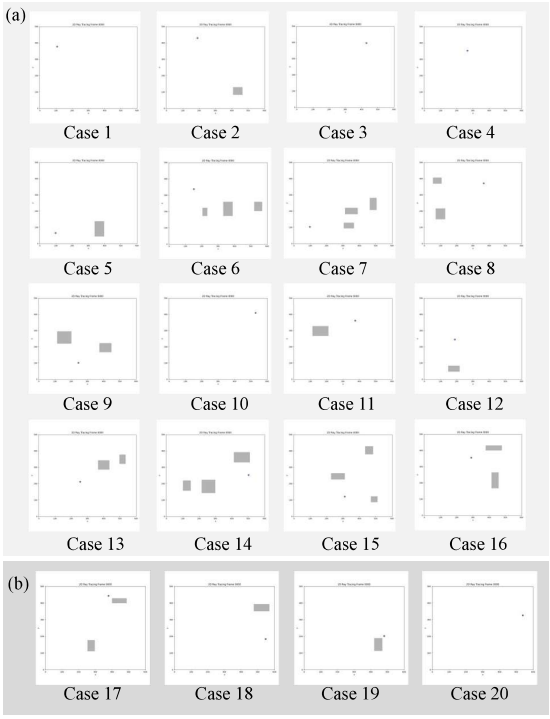


図 4: データセット．Case 1〜16：学習データ，Case 17〜20：テストデータ．

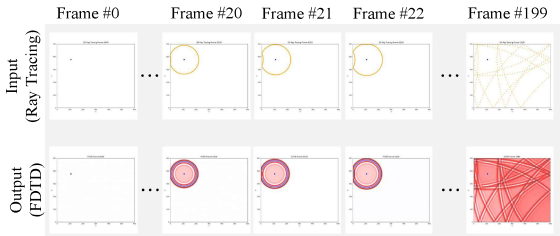


図 5: Case 1 における各フレームのペアの様子

4.2 実験結果と考察

図 6 に学習曲線を示す．Discriminator は不安定な状態が続いており，これは敵対的学習に典型的な挙動である．Generator の損失は一貫して減少したものの，明確な収束点には達しておらず，不完全な収束を示唆している．

図 7 に，学習済みモデルによって得られた波動予測結果を示す．図では，Case17〜Case20 の 4 つの解析条件について，Frame #40 および Frame #80 の結果を例として示している．左から順に，Ray Tracing による入力画像，機械学習モデルが生成した予測画像，Ground Truth として用いた FDTD の解析結果，そして生成画像の誤差マップを示している．

まず Case 20 について考える．Case 20 では障害物が存在せず，音源位置のみがトレーニングデータと異なる条件となっている．生成された波動場は Ground Truth

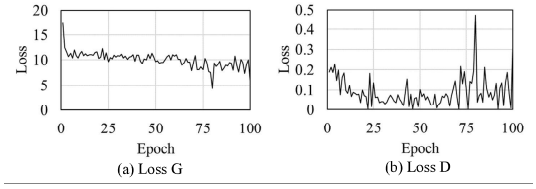


図 6: (a) Generator と (b) Discriminator の損失関数曲線．

とよく一致しており，良好な再現性が確認できた．

一方，Case 17〜19 では音源位置と障害物配置の両方が異なっている．全体的な波動パターンは類似していたものの，障害物による回折が再現されず，その結果として差異が生じた．

図 8 に，全フレームで平均した MSE(平均二乗誤差) を示す．障害物のない Case 20 では誤差は小さかったが，障害物を含む Case 17〜19 では誤差が大きくなった．これは，Ray Tracing が回折をモデル化せず，音線が直進のみを扱うため，入力画像には回折が反映されないことによる．その結果，音線情報の存在しない領域では予測が困難となった．しかしながら，音線が到達している領域においては，モデルは波動パターンを再現することができた．

今後，機械学習モデルの改善を行い，回折や干渉などの波動的な振る舞いをより正確に推定できるようにすることを試みる予定である．

5 おわりに

本研究では，image-to-image translation モデルを用いて，幾何音響から波動音響場を予測するフレームワーク Ray2Wave を提案した．2 次元の矩形モデルに対して，Ray Tracing と FDTD を実施し，それらを用いた機械学習を実施したところ，障害物がない場合や，音線が明示されている領域では波動伝搬の再現が見られた一方で，音線が表現しきれない障害物による回折については大きな誤差が生じる結果となった．

今後は機械学習モデルの改善に加え，広帯域や複数音源の条件へ拡張し，より大規模なシミュレーションデータセットの活用や，回折の扱いと汎化性能向上を目的とした物理インフォームドモデルの構築を目指す．

参考文献

[1] Shoichi Koyama, Tomoya Nishida, Keisuke Kimura, Takumi Abe, Natsuki Ueno, and Jesper Brunnström. MeshRIR: A dataset of room impulse responses on meshed grid points for

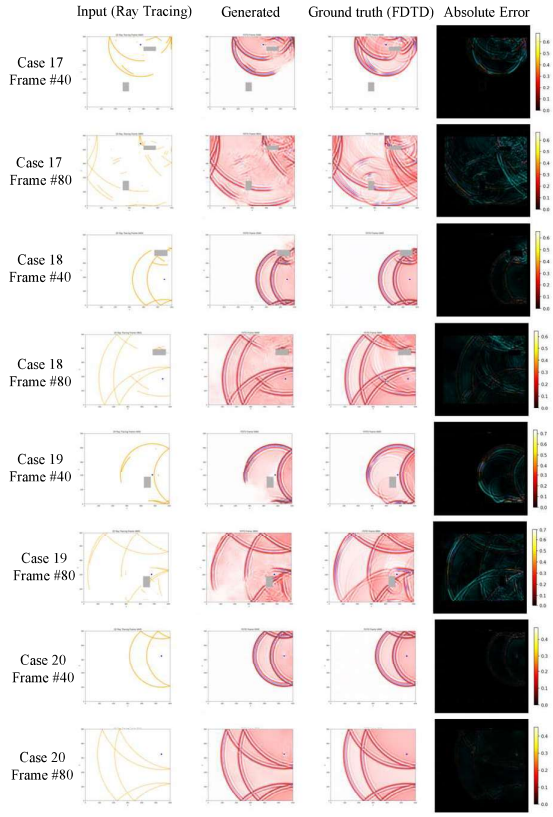


図 7: テストデータに対する予測結果と FDTD との比較（左から：GA 入力/予測結果/FDTD/誤差）.

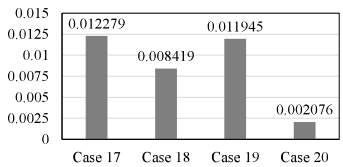


図 8: 各ケースに対する生成画像の MSE

evaluating sound field analysis and synthesis methods. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[2] Tony F. W. Embleton. Tutorial on sound propagation outdoors. *The Journal of the Acoustical Society of America*, 100(1):31–48, 1996.

[3] Shiguang Liu and Jin Liu. Outdoor sound propagation based on adaptive fdtd-pe. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 859–867, 2020.

[4] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *NeurIPS 2022 Datasets and Benchmarks Track*, 2022.

[5] Adil Alpkocak and Malik Kemal Sis. Computing impulse response of room acoustics using the ray-tracing method in time domain. *Archives of Acoustics*, 35(4):505–519, 2010.

[6] Fabian Brinkmann, Lukas Aspöck, David Ackermann, Steffen Lepa, Michael Vorländer, and Stefan Weinzierl. A round robin on room acoustical simulation and auralization. *The Journal of the Acoustical Society of America*, 145(4):2746–2760, 04 2019.

[7] Haibo Wang, Huimin Gao, and Ming Cai. Simulation of traffic noise both indoors and outdoors based on an integrated geometric acoustics method. *Building and Environment*, 160:106201, 2019.

[8] Takumi Yoshida, Takeshi Okuzono, and Kimihiro Sakagami. Binaural auralization of room acoustics with a highly scalable wave-based acoustics simulation. *Applied Sciences*, 13(2832), 2023.

[9] Takao Tsuchiya. Recent techniques on sound field simulation. *Japanese Journal of Applied Physics*, 61(SG0801), 2022.

[10] Brian Hamilton and Stefan Bilbao. FDTD methods for 3-D room acoustics simulation with high-order accuracy in space and time. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(11):2112–2124, 2017.

[11] Giulia Fratoni, Brian Hamilton, and Dario D'Orazio. Feasibility of a finite-difference time-domain model in large-scale acoustic simulations. *The Journal of the Acoustical Society of America*, 152(1):330–341, 07 2022.

[12] Masahiro Toyoda and Yuta Sakayoshi. Filter and correction for a hybrid sound field analysis of geometrical and wave-based acoustics. *Acoustical Science and Technology*, 42(4):170–180, 2021.

[13] Zhenyu Tang, Rohith Aralikatti, Anton Ratnarajah, and Dinesh Manocha. GWA: A LARGE

high-quality acoustic dataset for audio processing. In *Proceedings of the ACM SIGGRAPH Conference*, pages 1–9, 2022.

[14] Benjamin Moseley, Andrew Markham, and Tarje Nissen-Meyer. Fast approximate simulation of seismic waves with deep learning. *arXiv preprint arXiv:1807.06873*, 2018.

[15] Marco Olivieri, Fabio Antonacci, and Augusto Sarti. Near-field acoustic holography on arbitrary shapes using convolutional neural networks. In *Proc. INTER-NOISE 2021*, 2021.

[16] Sifan Wang, Shyam Sankaran, Hanwen Wang, and Paris Perdikaris. An expert's guide to training physics-informed neural networks. *arXiv preprint arXiv:2308.08468*, 2023.

[17] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *arXiv preprint arXiv:2006.09661*, 2020.

[18] He Ming Yao and Li Jun Jiang. Machine learning based neural network solving methods for the fdtd method. In *2018 IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting*, pages 2321–2322, 2018.

[19] H. M. Yao and W. Jiang. Machine learning based pml for the fdtd method. *Journal of Electromagnetic Waves and Applications*, 32(13):1729–1742, 2018.

[20] Li Siahkoohi, Mathias Louboutin, and Felix J. Herrmann. Neural network augmented wave-equation simulation. *arXiv preprint arXiv:1910.00925*, 2019.

[21] B. Moseley, A. Markham, and T. Nissen-Meyer. Solving the wave equation with physics-informed deep learning. *arXiv preprint arXiv:2006.11894*, 2020.

[22] Hao Wang, Jian Li, Linfeng Wang, Lin Liang, Zhoumo Zeng, and Yang Liu. On acoustic fields of complex scatters based on physics-informed neural networks. *Ultrasonics*, 128:106872, 2023.

[23] Shoichi Koyama, Juliano G. C. Ribeiro, Tomohiko Nakamura, Natsuki Ueno, and Mirco Pezzoli. Physics-informed machine learning for sound field estimation: Fundamentals, state of the art, and challenges. *IEEE Signal Processing Magazine*, 41(6):60–71, 2024.

[24] Marco Olivieri, Xenofon Karakonstantis, Mirco Pezzoli, Fabio Antonacci, Augusto Sarti, and Efren Fernandez-Grande. Physics-informed neural network for volumetric sound field reconstruction of speech signals. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(42):1–14, 2024.

[25] Juliano G. C. Ribeiro, Shoichi Koyama, Ryosuke Horiuchi, and Hiroshi Saruwatari. Sound field estimation based on physics-constrained kernel interpolation adapted to environment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4369–4382, 2024.

[26] Mirco Pezzoli, Fabio Antonacci, and Augusto Sarti. Implicit neural representation with physics-informed neural networks for the reconstruction of the early part of room impulse responses. In *10th Convention of the European Acoustics Association (Forum Acusticum)*, Turin, Italy, September 2023.

[27] Zitong Lan, Chenhao Zheng, Zhiwei Zheng, and Mingmin Zhao. Acoustic volume rendering for neural impulse response fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[28] Anton Ratnarajah, Zhenyu Tang, Rohith Aralikatti, and Dinesh Manocha. Mesh2ir: Neural acoustic impulse response generator for complex 3d scenes. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 924–933, New York, NY, USA, 2022. Association for Computing Machinery.

[29] Ziqi Fan, Vibhav Vineet, Hannes Gamper, and Nikunj Raghuvanshi. Fast acoustic scattering using convolutional neural networks. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 171–175, 2020.

[30] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *Proc. CVPR*, 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-068-09

話者情報を利用した距離ベース音声強調の改善

Improvement of Distance-based Speech Enhancement Using Speaker Information

糸山 克寿 [1*]　畑 和也 [2]　住田 直亮 [2]　中臺 一博 [1]
Katsutoshi Itoyama[1]　Kazuya Hata[2]　Naoaki Sumida[2]　Kazuhiro Nakadai[1]

[1] 東京科学大学　[2] 株式会社ホンダ・リサーチ・インスティチュート・ジャパン
[1] Institute of Science Tokyo　[2] Honda Research Institute Japan Co., Ltd.

Abstract: Abstract (English) comes here.....

1　はじめに

深層学習技術の進展により音声認識の性能は飛躍的に向上し、会議記録作成や音声アシスタントなど、多くの場面で実用化が進んでいる。これらのタスクでは、混合音の中から目的話者の発話を正確に抽出することが不可欠である。しかし一般的なデバイスに搭載されているのは単一マイクロホンであり、モノラル信号からは音源方向を推定するための位相差情報が得られない。そのため、複数話者が同時に存在する環境では、誰が発話したかを音響信号のみで適切に識別することができない。

近距離音声を強調する技術が実現できれば、たとえばスマートフォンに収録される音声は所有者の声のみとなり、会議などにおいて発話者と発話内容を自動的に対応付けられる。また、特別なマイクロホンアレイを必要とせず、普及率の高い一般デバイスだけで高精度な議事録作成や音声認識を実現できるという大きな利点がある。

本研究の目標は、**単一マイクロホンから得られるモノラル信号のみを用いて、マイク近傍の話者（主としてデバイス所有者）の音声を高精度に抽出すること**である。これにより、

- 会議など多人数環境でも、各デバイスに録音されるのはその所有者の音声のみとなり、発話内容と発話者を確実に対応付けられる
- スマートフォンやロボットといった日常的なデバイスの音声認識性能が、雑音や他者の発話の影響を受けにくくなる

といった効果が期待できる。

*連絡先：東京科学大学 工学院 システム制御系
〒 152-8552 東京都目黒区大岡山 2-12-1-W8-18
E-mail: itoyama@i.sc.eng.isct.ac.jp

この目的の達成に向けては、以下の技術的課題を解決する必要がある。

1. モノラル信号では方向情報が得られない単一マイクロホンでは位相差情報が得られないため、複数の近距離話者が存在した場合に誰の声かを識別できない。
2. 距離ベース手法 (DSE) の限界 DSE は直接音と反射音の比率を利用することで近距離音声を強調できるが、近距離に他者が存在すると誤って抽出してしまう、話者識別の機能を持たない、といった課題があり、単一マイクロホンでの確実な近距離音声抽出には不十分である。
3. 話者ベース手法 (TSE) の限界 TSE は話者特徴が事前登録された場合のみ有効であり、未登録話者の抽出ができない。そのため、近距離にある音声を抽出したいという今回の目的とは一致しない。

本稿では、上記の課題を踏まえて、単一マイクロホンを用いた近距離音声強調手法を提案する。提案手法では、

- 距離に依存して変化する音響的手がかり（直接音・反射音の比率、残響特性など）を利用する
- モノラル信号のみから、近距離音声を抽出し遠距離音を抑圧する
- 会議や日常環境において、デバイス所有者の音声のみを取得できる状況を実現する

といった点に重点を置く。提案手法は、特殊なマイクロホンアレイを必要とせず、一般的なスマートフォンなどに搭載された単一マイクロホンのみで利用可能である点が大きな特徴である。

2 関連研究

本節では、本研究に関連する研究について述べ、それらの課題を検討する。

2.1 時間周波数マスクによる音声強調

音声信号は時間とともに振幅が変化する一次元の波形として観測されるが、その性質をより詳細に扱うため、多くの音声強調手法では時間と周波数の両方の変化を表す時間周波数表現が用いられる。代表的な時間周波数表現がスペクトログラムであり、横軸を時間、縦軸を周波数、色の濃淡で各時間周波数成分の強度を表す。無音区間では全体的にエネルギーが低く、スペクトログラム上では暗い（あるいは青い）領域として表現される。

近距離音声と遠距離音声が混合している場合、それぞれは時間周波数平面上に異なるパターンとして現れる。例えば、ある区間では近距離話者の発話が強く現れ、その後に遠距離話者の発話が主となるような場合、混合音声のスペクトログラムは、近距離音声に対応する強い成分と、遠距離音声に対応する弱い成分が重畳した形となる。

時間周波数マスクは、この混合スペクトログラムに対して要素ごとに乗算されるフィルタであり、モノラル音響信号に対する音声強調手法として広く用いられている。混合信号のスペクトログラムを $Y(t,f)$、マスクを $M(t,f)$ とすると、強調後のスペクトログラム $\hat{S}(t,f)$ は

$$\hat{S}(t,f) = M(t,f) \odot Y(t,f) \tag{1}$$

のように要素積で表される。ここで、$M(t,f)$ は 0 から 1 の範囲の値であり、1 に近いほどその時間周波数成分をそのまま通過させ、0 に近いほど抑圧することを意味する。近距離音声強調の文脈では、「近距離音声に由来すると推定される成分のマスク値を 1 に近づけ、それ以外（遠距離音声や雑音）のマスク値を 0 に近づける」ことで、近距離音声のみから構成されるスペクトログラムを得ることを目指す。

近年の深層学習ベースの音声強調・音源分離手法では、この時間周波数マスクをニューラルネットワークによって直接推定する枠組みが主流となっている。本研究で扱う距離ベース音声強調（DSE）およびターゲット話者抽出（TSE）も、いずれも時間周波数マスキングの枠組みの上に構成されている。

2.2 距離ベース音声強調

距離ベース音声強調（Distance-based Speech Enhancement, DSE）は、音源とマイクロホンの距離に依存する音響的性質を利用して、近距離音声を強調する手法である。室内環境においてマイクで観測される音は、音源から直接届く「直接音」と、壁や天井等で反射した「残響音」の和として捉えることができる。直接音のエネルギーは音源との距離の二乗に反比例して減衰する一方、残響音は空間内で比較的均一なエネルギー分布を持ち、距離による減衰が緩やかである。その結果、近距離音源ほど Direct-to-Reverberant Ratio（DRR）が高くなる傾向がある。

DSE は、この DRR を主な手がかりとして、時間周波数マスクを用いて近距離成分を強調し、遠距離成分を抑制する。具体的には、混合音声のスペクトログラム上の各時間周波数ビンに対して、そのビンが近距離音声に由来するか遠距離音声に由来するかを推定し、近距離と推定されるビンのマスク値を高く、遠距離と推定されるビンのマスク値を低く設定する。

距離ベース音源分離という枠組みは、Patterson らによって初めて提案された。彼らは、近距離音声と遠距離音声を混合したモノラル信号のスペクトログラムを入力とし、畳み込みニューラルネットワーク（CNN）と再帰型ニューラルネットワーク（LSTM）を組み合わせたモデルにより、近距離音声を通過させる時間周波数マスクを推定する手法を提案した。この研究は、単一チャネル（モノラル）音声信号を用いて距離ベースの音声強調が可能であることを示した点で重要である。しかし、Patterson らの手法では、混合音声が近距離音声と遠距離音声のみから構成されており、それ以外の背景雑音が考慮されていない。また、仮想空間に配置した仮想マイクロホン・音源に基づくシミュレーション環境でのみ評価が行われており、実環境の録音に対する検証が十分ではなかった。

石井らは、Patterson らの手法を発展させ、より実環境に近い条件での距離ベース音声強調を実現した。具体的には、背景雑音を加えた上で、各音源に対して実環境で収録したインパルス応答を畳み込むことで学習データを生成し、Patterson らの CNN を表現力の高い ResNet に、LSTM を高速な GRU に置き換えたネットワークアーキテクチャを構築した。これにより、距離ベース音声強調の性能向上と計算効率の改善が報告されている。

一方で、これら DSE 系手法には依然として課題が残る。近距離・遠距離という距離情報のみに基づいてマスクを推定するため、近距離に複数の話者が存在する場合、それらを区別することができない。また、背景雑音の抑圧が過剰になった結果として音声成分が削られたり、逆に雑音が取り残されたりすることがあり、これらが後段の音声認識に悪影響を与えることが指摘されている。

2.3 ターゲット話者抽出

ターゲット話者抽出（Target Speaker Extraction, TSE）は、混合音声の中から特定の話者の音声のみを選択的に分離・強調する技術である。DSE が距離という物理的指標を利用するのに対し、TSE は話者固有の声質や話し方などの「話者性」を利用する点に特徴がある。

TSE では、分離対象としたい話者の参照音声を事前に用意し、その音声から抽出した話者埋め込みをネットワークへの条件情報として与える。モデルは、この話者埋め込みをもとに、混合音声のスペクトログラム上でターゲット話者に対応する時間周波数ビンを通過させるマスクを推定する。これにより、特定の人物の音声を高精度に抽出することが可能となる。

しかし、TSE もまた根本的な制約を持つ。参照音声によって指定された話者のみが抽出対象となるため、事前に音声を登録していない未知話者の音声を抽出することができない。会議への新規参加者や、不特定多数が存在する環境での利用を想定した場合、この柔軟性の欠如は大きな課題となる。

3 提案手法

本節では、混合音声スペクトログラムから近距離話者の音声スペクトログラムを推定するための、提案手法の構造について述べる。提案モデルは、ResNet ベースのエンコーダ、GRU を用いたデコーダ、および wav2vec 2.0 によって抽出された話者特徴量を用いる構成となっている。図 3.1 に提案アーキテクチャの概要を示す。

本モデルでは、観測された混合信号のスペクトログラムを $X \in \mathbb{C}^{T\times F}$ とし、これが近距離話者の音声スペクトログラム X_{near} と遠距離話者の音声スペクトログラム X_{far} から構成されるものと仮定する。

$$X = X_{\text{near}} + X_{\text{far}} \tag{2}$$

本研究の目的は、混合信号 X から近距離音声スペクトログラム X_{near} を推定することであり、時間周波数マスク推定に基づく音声強調を行う。

3.1 ResNet による混合音特徴抽出

本モデルのエンコーダ部には、深層残差ネットワーク（ResNet）を採用する。ResNet は残差接続による勾配消失の抑制を通じて深いネットワーク構造を可能にし、通常の CNN よりも高い表現能力を持つ。

音声スペクトログラムは時間 × 周波数の 2 次元構造を持つため、画像と同様に畳み込み処理が適用できる。本研究では、入力スペクトログラム Y を ResNet に入力し、時間・周波数構造を保持したまま高次特徴量 $H \in \mathbb{R}^{C\times T\times F}$ を抽出する。ここで、全ての畳み込み層のストライドを 1 に設定し、画像認識用途で一般的なダウンサンプリングを行わない。これにより、時間方向・周波数方向の解像度を保ったまま、後段のマスク推定に必要な詳細な特徴表現を得ることができる。

3.2 GRU によるマスク推定

デコーダには、再帰型ニューラルネットワークの一種である GRU（Gated Recurrent Unit）を用いる。GRU は LSTM と比較してゲート構造が簡潔であり、計算効率に優れる一方、長期依存関係の獲得能力はほぼ同等とされているため、本研究の音声強調タスクに適している。

ResNet で得られた特徴量マップ $H \in \mathbb{R}^{C\times T\times F}$ を周波数ビンごとに $H_f \in C \times T$ に分割し、各ビンに対して独立な GRU を適用する。これにより、周波数ごとに時間方向の依存関係をモデル化できる。

GRU の出力は全結合層（FC）を通じて時間周波数マスク値へと変換される。具体的には、次の処理が各周波数ビンに対して独立に行われる。

1. 全結合層（入力次元: 30 → 出力次元: 16）
2. ReLU
3. 全結合層（入力次元: 16 → 出力次元: 1）
4. ReLU（マスク値を非負に制限）

これにより、各時間フレームに対して 1 次元のマスク値が生成され、全周波数ビンに対してまとめることで最終的な時間周波数マスク $M \in \mathbb{R}^{T\times F}$ が得られる。

3.3 wav2vec 2.0 による話者特徴量抽出

提案手法では、話者識別に有効であることが示されている wav2vec 2.0 を用いて、近距離話者の特徴量を取得する。wav2vec 2.0 に参照音声を入力し、その出力特徴マップを時間方向に平均することで、固定長の話者ベクトル $F_{\text{near}} \in \mathbb{R}^S$ を得る。本研究では特徴量次元として $S = 256$ を採用する。

図 3.2 に示すように、wav2vec 2.0 により抽出された特徴量は話者ごとに分布が明確に分離しており、未学習話者に対しても適切な話者特徴量が得られることが確認されている。このように、wav2vec 2.0 は従来手法では利用されていなかった新たな構成要素であり、提案手法における話者特徴量抽出を担う重要なブロックである。

4 評価実験

本節では、提案手法の評価のための実験について述べる。実験の目的は、提案手法が従来手法よりも優れた近距離音声抽出性能を実現できるかどうか、また、抽出された近距離音声が正しく音声認識されるかどうかを検証することである。近距離音抽出性能の評価には、音声強調でよく用いられる Scale-Invariant Signal-to-Distortion Ratio improvement (SI-SDRi) を、音声認識性能の評価には、認識結果の文字誤り率 (Character Error Rate, CER) を、それぞれ指標として用いた。音声認識には、End-to-End 音声処理ツールキットである ESPnet を、日本語話し言葉コーパス (CSJ) で訓練したモデルを用いた。

本実験では、話者特徴を用いる提案手法と、話者特徴を用いない石井らの距離ベース音声強調手法を比較した。石井らの手法は、提案手法から話者特徴抽出部分と FiLM による特徴統合部分を取り除いたものに相当する。

4.1 データセット

本実験における各手法の訓練および評価のために構築した、摸擬対話音声データセットについて述べる。

摸擬対話音声は、近距離音声、遠距離音声、背景雑音の 3 つの信号の組からなる。二人の話者が対話している状況を想定し、近距離音声と遠距離音声は、それぞれの話者が交互に発話するように構成する。実際の対話ではしばしば発話のオーバーラップが生じるため、オーバーラップ割合を定め、対話全体の長さに対してその割合の区間が、近距離音声と遠距離音声がオーバーラップするように構成する。

具体的には以下の手順で近距離音声と遠距離音声をそれぞれ構築した。

1. 無残響環境で収録された音声コーパスから、それぞれ異なる話者の音声を 2 つ抽出する。
2. 各音声を、音声コーパスに付与された区間情報に基づいて、数秒程度の発話区間に分割する。
3. 時間軸に各話者の発話区間を交互に配置し、対話音声を構築する。この際、事前に定めたオーバーラップ割合に基づいて、発話の一部がオーバーラップするように、区間の開始・終了時刻を調整する。

背景雑音は、事前に用意した背景雑音データベースからランダムに選択したものを組み合わせて用いた。

近距離音声と遠距離音声の作成に用いたインパルス応答について述べる。会議や打ち合わせでの使用を想定した評価を行うため、大きさが 3m x 6m、幅 150 cm の会議机が 5 台設置された会議室でインパルス応答を収録した。一般的なオフィス環境であり、防音設備はないものの、室内にはエアコンや OA 機器等の雑音源が存在する。

この会議室で、2 名が別々の机に着席して打ち合わせを行っている場面を想定した。この設定を模倣するため、以下のように機材を配置した。

- マイクロホン（スマートフォン）
  - 使用機種：iPhone XR, iPhone SE2, iPhone SE3 のいずれか
  - 机上、ノート PC の左側に設置
- スピーカー
  - 使用機種：Genelec 8020D スタジオ・モニター
  - 机上 40cm の高さ、机の手前側にスタンドで設置
- ノート PC
  - 机上、スピーカーの正面に設置
  - 現実の会議における音の反射・回折の効果を再現する目的で、ディスプレイを開いた状態と閉じた状態に設定

この条件に基づき、合計で 640 個のインパルス応答を収録した。

混合音声は以下のように生成した。

1. オーバーラップ割合に基づいて構成した 2 つの対話音声に対して、一方に近距離インパルス応答を、もう一方に遠距離インパルス応答を畳み込み、近距離音声と遠距離音声を作成する。この近距離音声が出力信号となる。
2. 雑音データベースから雑音を抽出し、近距離音声、遠距離音声、雑音を加算した混合音を作成する。この混合音が入力信号となる。
3. 近距離音声を話者特徴抽出器に入力し、話者特徴ベクトルを生成する。

上記の手順に基づいて、訓練データセットと評価データセットを作成した。摸擬対話音声作成のためのデータセットには日本語話し言葉コーパス CSJ を用いた。訓練データセットと評価データセットの分離のため、eval2 サブセットを評価データセットに、それ以外のサブセットを訓練データセットに用いた。畳み込むインパルス応答も同様に、訓練データセット用と評価データセット用のものが重複しないように分割した。訓練

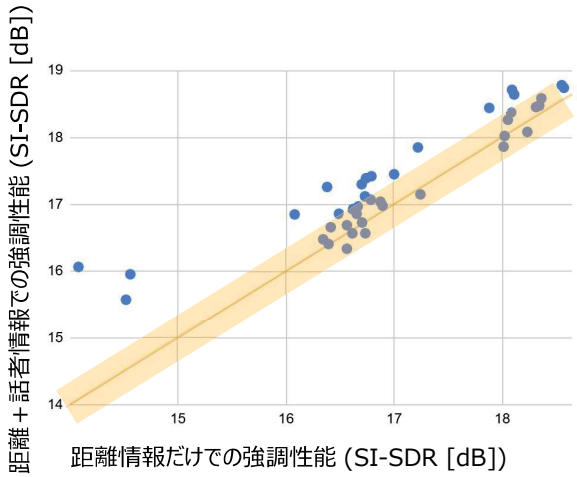


図 1: 音声強調性能評価結果。各点は、各評価データを提案手法と比較手法とで強調した結果の SI-SDR を示す。対角線は提案手法と比較手法の強調性能が同等であるような領域であり、この対角線よりも上方向に点が存在すれば、提案手法で音声強調性能が向上したことを示している。

データセットとして長さ 30 秒の摸擬対話音声を 1000 サンプル、評価データセットとして長さ 5 分の摸擬対話音声を 5 サンプル生成して用いた。

4.2 実験結果

図 1 に音声強調性能評価結果を示す。各点は、各評価データを提案手法と比較手法とで強調した結果の SI-SDR を示す。図中の対角線は、提案手法の音声強調性能（縦軸）と比較手法の強調性能（横軸）とが同等であるような領域であり、この対角線よりも上方向に点が存在すれば、提案手法で音声強調性能が向上したことを示している。すべての評価データが、対角線よりも上の領域にプロットされていることから、話者特徴の導入は音声強調性能の改善に寄与することが分かる。図の左側に対角線より上方にプロットされた点がいくつかあることから、話者情報なしでは音声強調性能が低いような評価データに対しては、話者情報の導入により大きく性能向上が可能であることが示唆される。図 2 に音声認識性能評価結果を示す。強調性能の評価結果と同様に、話者特徴の導入により、音声強調性能も改善していることが分かる。

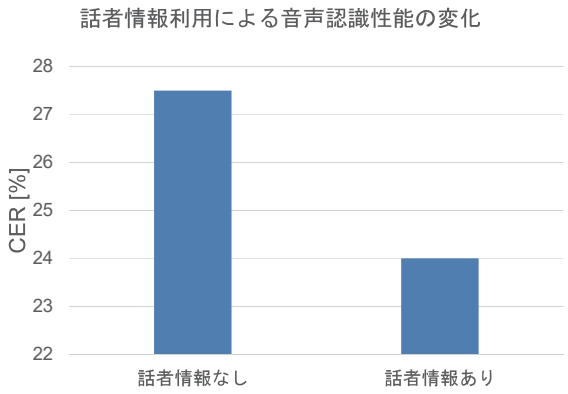


図 2: 音声認識性能評価結果

5 まとめ

謝辞

□□□....

参考文献