# 空間情報を伴う音響言語モデルの検討 *

◎関健太郎（東大／慶大），岡本悠希，山岡洸瑛，齋藤佑樹（東大），
高道慎之介（東大／慶大），猿渡洋（東大）

## 1　はじめに

近年，音響と言語のマルチモーダル学習において，音響信号とキャプションを結びつける対照言語音響事前学習（contrastive language-audio pretraining: CLAP） [1] が注目されている．この手法は音響信号と自然言語による音響信号の説明文であるキャプションのペアデータを用いた対照学習によって両モーダルを結びつけた埋め込み表現を獲得しており，自然言語を用いた音響信号のゼロショット分類 [1]，音響信号から対応するキャプションを生成する自動音響キャプショニング [2]，さらには自然言語クエリを用いた音源分離 [3] など，様々なタスクに応用されている．

しかし，従来の CLAP は音源情報，すなわちどのような音であるのかという情報（例：犬の鳴き声，拍手など）に注目しており，音が「どこで鳴っているのか」といった空間情報は取り入れられていない．一方で，例えば人間は音の種類だけでなく方向などを把握することにより，高度な音環境認識を行っている．このように，実世界の音響信号処理タスクにおいて空間情報は重要となる．

そこで本研究では，空間情報を伴う音響信号を扱う空間拡張型 CLAP を提案する．提案する空間拡張型 CLAP モデルにおける音響情報エンコーダは，音源情報を表現するための音源情報エンコーダと，音源情報に空間情報を結びつけるための空間情報エンコーダの 2 つから構成される．空間情報エンコーダは複数の音源の空間情報を同時に捉えるために，ポリフォニック SELD（sound event localization and detection）のタスクで事前学習される．さらに，複数音源が存在する状況において音源情報と空間情報の正しい対応を学習させるためのデータ拡張手法として，空間対照学習を提案する．

実験的評価により，空間拡張型 CLAP の特性を評価する．評価は音源情報，空間情報，および音源情報と空間情報の結びつきの 3 つの観点から実施する．従来手法との比較により，提案手法の有効性を確認する．

## 2　提案手法

本研究では複数音源を扱う空間拡張型 CLAP と，そのための学習手法である空間対照学習を提案する．

### 2.1　音響情報エンコーダ

先行研究 [4] では，音源情報エンコーダと空間情報エンコーダの 2 つの並列なエンコーダの出力を multi-layer perceptron (MLP) で統合することにより空間情報を伴う音響信号表現を獲得している．同様に，本研究でも Fig. 1 に示すように，音源情報エンコーダ

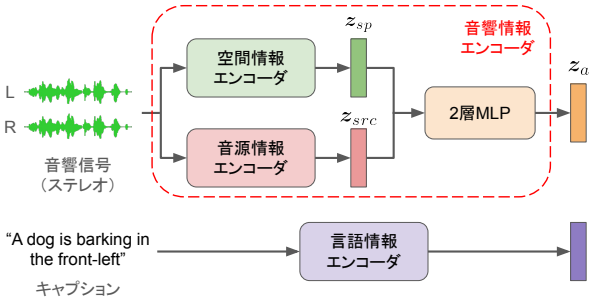


Fig. 1: 空間拡張型 CLAP の全体像．

と空間情報エンコーダの 2 つのエンコーダ出力を 2 層 MLP で統合し，その出力を音響情報エンコーダの出力とする．

CLAP の学習では音とキャプションのペアデータが必要となるが，大量のペアデータを用意することは難しいため，各モーダルで事前学習されたモデルをエンコーダとして用いる．そこで，音源情報エンコーダにはモノラル信号を対象とする CLAP の学習済み音響情報エンコーダを用いる．

空間情報エンコーダについても同様に，事前学習済みモデルを用いる．先行研究 [4] では到来角（direction of arrival: DoA）推定のタスクで事前学習したモデルを空間情報エンコーダとして利用しているが，この方法は音源情報と空間情報を別々に捉えているため，複数音源が存在している状況において音源情報と空間情報のアラインメントを取ることが困難であると予想される．そこで本研究では，複数の音響イベントが同時に発生するポリフォニック SELD のタスクにおいて用いられる枠組み [5–7] を採用した事前学習により，音源情報と空間情報の結びついた埋め込みベクトル $\boldsymbol{z}_{sp}$ を得る．

音源情報エンコーダの出力 $\boldsymbol{z}_{src}$ と空間情報エンコーダの出力 $\boldsymbol{z}_{sp}$ は concatenate され，2 層 MLP によって統合される．本研究における空間情報エンコーダは SELD のタスクであり音源情報を扱うため，空間情報エンコーダのみで両者の情報を表現することも可能である．しかし実際には音源情報は空間情報より多様かつ詳細な記述に対応しているため，音源情報のエンコードに特化した音源情報エンコーダの情報との統合によってキャプション埋め込みとの類似度を高めることを狙う．

### 2.2　言語情報エンコーダ

従来のモノラル信号を対象とした CLAP [1] と同様，大規模なテキストで事前学習された言語情報エンコーダモデルを用いる．膨大なドメインで事前学習された言語情報エンコーダを音源情報のキャプションという限定的なドメインでファインチューニングすること

* Study on Spatially-Aware Audio-Language Model, by Kentaro Seki (The University of Tokyo/Keio University), Yuki Okamoto, Kouei Yamaoka, Yuki Saito (The University of Tokyo), Shinnosuke Takamichi (The University of Tokyo/Keio University), Hiroshi Saruwatari (The University of Tokyo).

で，このドメインに特化した，かつ音響信号の表現と結びついた表現を獲得することを狙う．

### 2.3 学習方法

#### 2.3.1 データの作成

従来の CLAP [1] は AudioCaps [8] や Clotho [9] といったデータセットを用いて学習されていたが，これらの音響信号はモノラルであり空間情報が含まれていない．そこで本研究では，室内インパルス応答（room impulse response: RIR）のデータセットを用いてマルチチャネル信号と自然キャプションのペアデータセットを作成する．

**単一音源の場合**：モノラル音響信号 $x(t)$ と RIR $\boldsymbol{h}(t)$ の畳み込みによりマルチチャネル音響信号を作成する：

$$\boldsymbol{y}(t) = (\boldsymbol{h} * x)(t) := \int_{-\infty}^{\infty} \boldsymbol{h}(\tau)\, x(t-\tau)\, d\tau. \quad (1)$$

また，$\boldsymbol{y}(t)$ に対応するキャプション $D$ については，DoA $\theta$ に基づいてその空間的な情報を記述するキャプション $S_\theta$ を生成し，$x(t)$ に対応するキャプション $C$ の文末に $S_\theta$ を追記することによって作成する．$\theta$ に基づいて $S_\theta$ を生成する方法として，ここでは角度の範囲を有限クラスに分割し，それぞれの記述を事前に割り当てる方法を取る．例えば，$C$ が “A dog is barking.” であり，$S_\theta$ が “in the right side” であるとき，$D$ は “A dog is barking in the right side.” となる．

**複数音源の場合**：複数のモノラル音響信号 $\{(x_i(t)\}_{i=1}^n$ と同数の RIR $\{\boldsymbol{h}_i(t)\}_{i=1}^n$ を用いて，マルチチャネル信号 $\boldsymbol{y}(t)$ を以下のように作成する：

$$\boldsymbol{y}(t) = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{h}_i * x_i)(t). \quad (2)$$

$\boldsymbol{y}(t)$ は $x_i(t)$ と $h_i(t)$ で作成されるマルチチャネル信号 $\boldsymbol{y}_i(t)$ を足し合わせたものである．また，$\boldsymbol{y}(t)$ に対応するキャプション $D$ は，$\boldsymbol{y}_i(t)$ に対応するキャプション $D_i$ を結合することで作成する．

#### 2.3.2 空間対照学習

本研究では「空間対照学習」という学習方法を提案する．この方法は音源情報と空間情報の組み合わせを入れ替えたサンプルを作成することによって，複数音源の状況において音源情報と空間情報が結びついた表現を獲得することを狙う．

空間対照学習は次のように定式化される．$n$ 個の複数のモノラル音響信号 $\{(x_i(t)\}_{i=1}^n$ と同数の RIR $\{\boldsymbol{h}_i(t)\}_{i=1}^n$ を用いて，$n!$ 個のマルチチャネル信号 $\boldsymbol{y}_j(t)$ を作成する：

$$\boldsymbol{y}_j(t) = \frac{1}{n}\sum_{i=1}^{n} \left(\boldsymbol{h}_{\sigma_j(i)} * x_i\right)(t). \quad (3)$$

ただし $\sigma_j$ は $n$ 個のインデックスに対する $j$ 番目の置換を表し，$\{\sigma_j\}_{j=1}^{n!}$ は $n$ 次の置換の集合を表す．例えば $n=2$ の場合，次の 2 つのサンプルを作成する：

$$\boldsymbol{y}_1(t) = (\boldsymbol{h}_1 * x_1)(t) + (\boldsymbol{h}_2 * x_2)(t). \quad (4)$$

$$\boldsymbol{y}_2(t) = (\boldsymbol{h}_2 * x_1)(t) + (\boldsymbol{h}_1 * x_2)(t). \quad (5)$$

$\{\boldsymbol{y}_j(t)\}_{j=1}^{n!}$ は音源情報のみで区別することができないため，これを用いた対照学習によって音源情報と空間情報の結びついた表現の獲得を狙う．

## 3 実験

本研究では空間拡張型 CLAP における適切な学習方法を調査するために，空間情報エンコーダの事前学習方法と対照学習における空間対照学習の有無による性能の違いを調査した．

### 3.1 実験条件

#### 3.1.1 データセット

モノラル音響信号とキャプションのペアデータセットとして，AudioCaps 2.0 [8] を用いた．このデータセットは YouTube 上の環境音を集めたデータセットである AudioSet [10] から派生したデータセットであり，AudioSet の学習サブセットの一部に人間がキャプションを付与したものである．

RIR と DoA のペアデータセットは pyroomacoustics [11] を用いたシミュレーションにより作成した．直方体の部屋の中心に 0.15 m 間隔でステレオマイクを設置した．部屋はサイズ ($x$ m, $y$ m, $z$ m) と吸収係数 $r$ によってパラメータ表現されており，$x, y, z$ はそれぞれ $[7.0, 10.5]$, $[6.0, 8.5]$, $[2.5, 4.5]$ の範囲から 0.5 刻みで値を選択し，$r$ は 0.4, 0.5, 0.6 の 3 通りから選んだ．$(x, y, z, r)$ の組として合計 440 の部屋が作成され，これをランダムに分割して学習・検証・評価サブセットを作成した．各サブセットのサイズは 360, 60, 60 である．音源の位置はマイク中心から 2 m とし，音源方向は一周を 360 分割して離散化したものから選択した．音速は 340 m/s とし，サンプリング周波数は 16 kHz とした．ステレオマイクでは前後が区別できないため，DoA ラベルは前後を折り返し，180° の範囲でラベルを付与した．DoA のキャプションは 180° を 5 等分し，“on the left side”, “in the front-left”, “in front”, “in the front-right”, “on the right side” の 5 通りを割り当てた．

モノラル音響信号 1 つにつき 1 つの RIR を割り当てて利用した．学習・検証サブセットについては組み合わせをエポックごとにランダムに変更し，評価サブセットについてはあらかじめ固定した 1 つを割り当てた．

#### 3.1.2 モデル構造

**音源情報エンコーダ**：先行研究 [1] において CLAP の枠組みに基づいて事前学習された HTS-AT[1] を用いた．このモデルは大量のデータを用いて事前学習されているため，音源情報のきめ細かな表現を得ることを期待する．

**空間情報エンコーダ**：SELDNet [5] に基づく構造を用いた．SELDNet は入力特徴量として振幅スペクトログラムと位相スペクトログラムを結合したテンソルを用い，これに 3 層の畳み込みニューラルネットワーク，2 層の双方向 Gated Recurrent Unit を適用する

[1] https://huggingface.co/lukewys/laion_clap/blob/main/music_speech_audioset_epoch_15_esc_89.98.pt

ことで中間特徴量を抽出し，2 つの 2 層 MLP によって音響イベント検出（sound event detection: SED）および DoA 推定を同時に実行する．本研究ではこの中間特徴量を時間軸に沿って平均することで固定次元の特徴量を計算し，これを $\boldsymbol{z}_{sp}$ として用いた．入力特徴量の計算における短時間フーリエ変換のパラメタは窓長を 1,024，シフト長を 512 とした．

**言語情報エンコーダ**：先行研究 [1] と同様に，RoBERTa の事前学習済みモデル[2]を用いた．

#### 3.1.3 比較する学習手法

学習手法の違いによるモデル性能の違いを評価する．本研究では空間情報エンコーダの事前学習と対照学習の 2 つについて調査した．

空間情報エンコーダの事前学習手法による性能の違いを評価するために，以下の手法を評価した．

**なし**：提案手法を従来のモノラル CLAP と比較するために，空間情報エンコーダを用いず，音源情報エンコーダの出力に MLP を適用したものを音響情報エンコーダ出力とした．

**DoA**：先行研究 [4] と同様に，DoA 推定によって事前学習を行った．Section 3.1.2 で述べた空間情報エンコーダの出力特徴量に線形層，ReLU 関数，線形層からなる 2 層 MLP を適用し，DoA を出力するよう学習した．本研究はステレオ信号を扱うため，DoA は角度 $\theta$ という 1 次元の量で扱い，損失関数として二乗誤差を用いた．この方法は SELD と異なり音源情報をエンコードする必要がないため，より精密に空間情報がエンコードされることが期待される．学習データとして，Section 3.1.1 で述べたデータセットに対し，Section 2.3.1 で述べた方法を用いて単一音源のマルチチャネル信号を作成した．

**SELD**：Section 3.1.2 で述べた空間情報エンコーダの出力特徴量に 2 つの MLP を適用し，それぞれ SED，DoA 推定を行うように学習した．DoA は DoA 推定の場合と同様に，方向 $\theta$ として出力された．損失関数はイベントクラスごとに定義され，SED，DoA はそれぞれバイナリクロスエントロピー，二乗誤差を損失とした．イベントクラスは AudioSet [10] において付与されたイベントクラスのラベルを利用した．学習データには Section 3.1.1 で述べたデータセットに対し，バッチの半分を単一音源として，残り半分は 2 音源からなる複数音源としたものを利用した．

バッチサイズは “DoA” において 64，“SELD+空間対照学習なし” において 48 とした．いずれの手法においても 1 バッチの作成に用いられるモノラル音響信号の個数は 64 であり，またエポック数は 50 とした．

対照学習では，従来の CLAP [1] と同様に，対照損失を用いて学習を行った．バッチサイズは 24 とし，1 バッチにつき 32 のモノラル音響信号を用いてマルチチャネル信号を作成した．“SELD+空間対照学習なし” と同様に，バッチ内の半数を単一音源として，残り半分を 2 音源からなる複数音源として用いた．また，空間事前学習を適用する場合は Section 2.3.2 で説明した方法によるデータ拡張によってバッチサイズが 32 となった．エポック数は 50 とした．

[2] https://huggingface.co/FacebookAI/roberta-base

### 3.2 評価指標

#### 3.2.1 R@1 score

音源情報の表現力を検証するため、音響信号とキャプションの相互検索（audio-to-text: A2T, text-to-audio: T2A）タスクを実施した。従来の CLAP モデルと提案する空間拡張型 CLAP とを Recall@1（R@1）の指標において比較評価した．キャプションには空間情報の記述を含む “sp+src” と，音源情報のみの記述である “src” の 2 通りで実験を行った．いずれの場合においてもキャプション候補はモノラル音響信号と同数である．

#### 3.2.2 DoA 分類

空間拡張型 CLAP における空間情報の表現精度を評価するため，単一音源のマルチチャネル信号に対する DoA クラスの分類精度を評価した．マルチチャネル信号の音響情報埋め込みに対して DoA キャプションの言語情報埋め込みとの類似度を比較し，最も類似度の高いキャプションが DoA に対応するキャプションである割合を評価した．

#### 3.2.3 音源-DoA 割り当て

空間拡張型 CLAP における音源情報と空間情報との結びつきを評価するために，以下の手順を実施する．まず，2 音源の混合信号と，各音源に対応するキャプションおよび DoA のキャプションが与えられる．このとき，音響情報埋め込みと言語情報埋め込みの類似度に基づいて，どの音源がどの DoA に対応するかを推定する．具体的には，全ての音源-DoA の組み合わせについてキャプションを作成して言語情報埋め込みと音響情報埋め込みとの類似度を算出し，最も高いスコアを示す組み合わせを対応関係として選択する．この評価により，空間拡張型 CLAP が音源情報と空間情報を個別に処理しているのではなく，両者を統合的に捉えているかどうかを検証する．

#### 3.2.4 類似度

音響情報の埋め込み空間と言語情報の埋め込み空間の類似度を調査するために，各ペアについて，音響情報エンコーダによる埋め込みと言語情報エンコーダによる埋め込みのコサイン類似度を評価した．

### 3.3 実験結果

各評価項目の結果を Table 1 に示す．

“DoA” と “SELD+空間対照学習なし” の比較においては，“DoA 分類” および “音源-DoA 割り当て” の性能はほぼ同等であったが，“DoA” は “R@1 (T: src)” において大幅な性能低下が見られた．さらに，空間記述を伴わない場合の埋め込み類似度も大きく低下しており，“DoA” では音響情報エンコーダの表現能力が空間情報の表現に過剰に割かれていた可能性がある．これは，複数音源が存在する状況下において，DoA エンコーダと音源情報エンコーダが完全に異なる情報を表現しているため，統合が不十分であったことを示唆している．このことから，SELD タスクによる事前学習が空間拡張型 CLAP において効果的であることが示された．

Table 1: 空間拡張型 CLAP モデルの比較．各評価指標ごとに最も高い数値を太字で示している．Chance rate は乱数出力における期待値の理論値である．「空間情報」は空間情報エンコーダの事前学習手法であり，「なし」は空間情報エンコーダを使わない手法である．

| 学習方法 | | 評価結果 | | | | | | 類似度 | |
|---|---|---|---|---|---|---|---|---|---|
| 空間情報 | 空間対照学習 | R@1 (T: sp+src) A2T | R@1 (T: sp+src) T2A | R@1 (T: src) A2T | R@1 (T: src) T2A | DoA 分類 | 音源-DoA 割り当て | 空間記述 あり | 空間記述 なし |
| (Chance rate) | | (0.01%) | (0.01%) | (0.01%) | (0.01%) | (20%) | (50%) | (0.0) | (0.0) |
| なし | X | 20.47% | 21.30% | 20.58% | **21.09%** | 17.80% | 49.79% | 0.795 | **0.798** |
| DoA | X | 41.05% | 41.46% | 13.68% | 9.57% | 86.21% | 70.16% | **0.881** | 0.553 |
| SELD | X | **44.14%** | **43.31%** | **21.71%** | 15.43% | 93.72% | 68.31% | 0.850 | 0.717 |
| SELD | O | 43.11% | 42.70% | 19.14% | 13.58% | **94.86%** | **79.01%** | 0.856 | 0.707 |

“SELD+空間対照学習なし” と “SELD+空間対照学習あり” の比較では，“R@1” および “DoA 分類” における大きな差は見られなかったが，“音源-DoA 割り当て” タスクにおいては後者が大きく性能を向上させた．これは空間対照学習によって音源情報と空間情報を結合した埋め込み表現がより効果的に学習されたことを示している．

一方，従来のモノラル CLAP に相当する “なし” と，空間拡張型 CLAP（“SELD+空間対照学習なし” および “SELD+空間対照学習あり”）を比較すると，空間拡張型 CLAP では “R@1 (T: src)” の性能が低下した．これは，空間情報を伴う CLAP においてはバッチ内のサンプルを空間情報によって識別できるために，音源情報に基づく識別が上手く学習されなかったものと考えられる．

### 3.4 埋め込み空間の分析

Fig. 2 に “SELD+空間対照学習なし” 手法における埋め込みベクトルを主成分分析によって可視化した結果を示す．“Audio” は音響情報の埋め込みベクトルであり，“Text: src+sp”，“Text: src”，“Text: sp” はそれぞれ空間情報と音源情報，音源情報のみ，空間情報のみを記述したキャプションの埋め込みベクトルを表す．

“Audio” の埋め込みはクラスごとにクラスタを形成しているが，クラスタ同士は分離されていない．これは DoA が連続量であり，クラスの境界付近において埋め込みベクトルが接近しているものと考えられる．“Text: src+sp” の埋め込みもクラスごとにクラスタを形成しており，“Audio” と比較してクラスタ間の距離が離れている．これは空間情報のキャプションが DoA を離散化したものに基づいているためと考えられる．

“Audio”，“Text: src+sp”，および “Text: sp” において各クラスの分布する領域は共通しており，空間情報が音響モーダル・言語モーダルに共通する表現を獲得していることがわかる．一方で “Text: src” の分布領域は限定的であり，この表現力低下によって音源情報を用いた検索タスクの性能が生じているものと考えられる．

## 4 まとめ

本研究では複数音源が存在する状況を扱う空間拡張型 CLAP を提案した．空間拡張型 CLAP の音響情報エンコーダは，音源情報エンコーダと空間情報エンコーダの情報を統合することによって音源ごとに空間情報が結びついた表現を獲得する．さらに，空間拡張型 CLAP のための学習手法である空間対照学習を提案した．実験的評価により空間拡張型 CLAP が空間情報の表現，音源情報と空間情報の結びつきを両立することを示し，特に空間対照学習が複数音源の状況を識別する上で有効であることを示した．一方で，空間拡張型 CLAP は音源情報の扱いにおいて性能が低下することが確認され，これは今後の課題である．

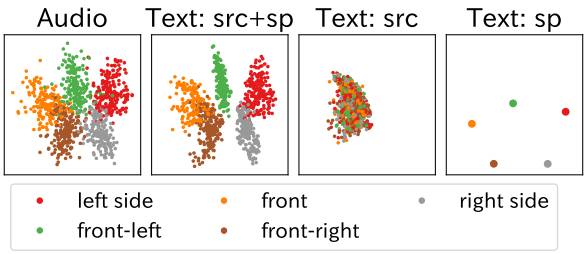


Fig. 2: 主成分分析を用いた埋め込み空間の可視化．

**謝辞:** 本研究は科研費 24KJ0860（アルゴリズム開発），JST ムーンショット型研究開発事業 JPMJMS2011（モデル学習），2025 年度国立情報学研究所公募型共同研究（251S4-22735）（データセット作成）および創発的研究支援事業 JPMJFR226V（評価実験）の助成を受け実施した．

## 参考文献

[1] Yusong Wu *et al.*, in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[2] Xiquan Li *et al.*, in *Proc. ICASSP*. IEEE, 2025, pp. 1–5.

[3] Xubo Liu *et al.*, in *Proc. Interspeech*, 2022, pp. 1801–1805.

[4] Bhavika Devnani *et al.*, *Proc. NeurIPS*, vol. 37, pp. 33 505–33 537, 2024.

[5] Sharath Adavanne *et al.*, *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.

[6] Kazuki Shimada *et al.*, in *Proc. ICASSP*. IEEE, 2021, pp. 915–919.

[7] Yin Cao *et al.*, in *Proc. ICASSP*. IEEE, 2021, pp. 885–889.

[8] Chris Dongjoo Kim *et al.*, in *NAACL-HLT*, 2019.

[9] Konstantinos Drossos *et al.*, in *Proc. ICASSP*. IEEE, 2020, pp. 736–740.

[10] Jort F Gemmeke *et al.*, in *Proc. ICASSP*. IEEE, 2017, pp. 776–780.

[11] Robin Scheibler *et al.*, in *Proc. ICASSP*. IEEE, 2018, pp. 351–355.