Spatial Audio Captioning: 複数音源状況下における空間情報を伴う説明文の生成とその評価*

◎関 健太郎 (東大/慶應大)，岡本 悠希 (東大)，山岡 洸瑛 (東大)，
齋藤 佑樹 (東大/産総研)，高道 慎之介 (慶應大/東大)，猿渡 洋 (東大)

1 はじめに

音響キャプショニング（Automated Audio Captioning）とは，入力された音響信号に対してその内容を記述する説明文（キャプション）を生成するタスクである．離散ラベルを用いた音響シーン分類と比べて，自然言語の持つ自由度と柔軟性を活用した詳細な情報の取得が可能であると期待されている．従来の音響キャプショニングは「何が鳴っているか」という音源情報の記述に焦点を当てていたが，音環境理解においては「どこで鳴っているか」という空間的情報も重要な情報である．そこで本研究では，音源情報と空間情報の双方を記述する空間的音響キャプショニング（Spatial Audio Captioning）を扱う（Fig. 1）．

空間的音響キャプショニングの課題は，複数の音源が存在する状況の扱いである．この状況では音源ごとに音源情報と空間情報を対応させて記述する必要があるため，単一の音源のみが存在する場合と異なり，音源情報と空間情報は同時に扱う必要が生じる．Fig. 2a に示すように音源情報と空間情報を独立に扱う音響エンコーダ [1] を用いると，音源情報と空間情報の対応関係に複数の可能性が生じ，適切な対応関係を記述することが困難となる．また，空間的音響キャプショニングの評価において，この対応関係の正確性の評価が課題となる．従来の音響キャプショニングでは SentenseBERT [2] や BERTScore [3] といった評価指標によって文意の類似度が評価される [4] が，これらは全体的な文意の類似度を測るものであるため，対応関係のような細かな差異の評価には適さないと考えられる．

そこで本研究では，複数音源状況下における空間的音響キャプショニングの手法と，その評価方法を検討する．提案手法は Spatial-CLAP [5] を音響エンコーダとして用いることで音源情報と空間情報の対応関係を認識し，両者を適切に統合したキャプションを生成する．また，生成キャプションの評価手法として，Spatial-CLAP [5] の言語エンコーダ埋め込みを主成分分析で得られた部分空間に射影しコサイン類似度を評価する手法を提案する．この手法は部分空間の構成法を制御することで音源情報・空間情報のいずれか一方に焦点を当てた評価が可能であり，空間的音響キャプショニングの多角的な評価手法を実現する．

実験では，参照キャプションと人工的に作成した誤りを含むキャプションとの類似度を評価することで既存の評価指標が特定の誤りの検出を困難とすることを示し，提案する評価手法によって多角的に評価できることを示す．この結果に基づいて，実際のキャプション生成手法を比較することにより，提案する空間的音響キャプショニング手法の有効性を確認する．

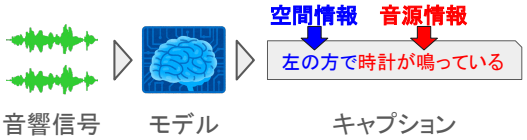
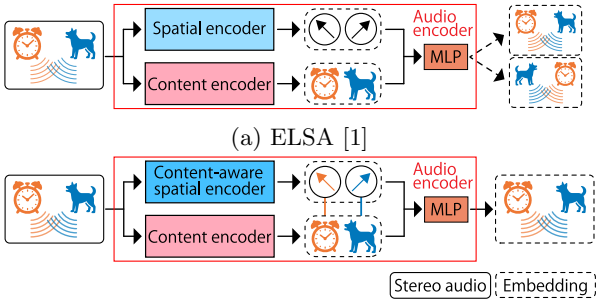


Fig. 1: 空間的音響キャプショニングの概念図．



(a) ELSA [1]

(b) Spatial-CLAP [5]

Fig. 2: 音源情報と空間情報を独立に扱う ELSA [1] による埋め込み表現は複数音源状況において音源と空間の対応を表現できないが，Spatial-CLAP [5] はこの状況下でも適切な埋め込み表現を学習する．

2 関連研究

2.1 音響キャプショニング

音響キャプショニングは，音響エンコーダによって抽出された音響特徴量を用いて言語デコーダを条件付けする encoder–decoder 構造によって実現される [4]．従来の音響エンコーダはモノラル信号を入力とし，音源イベントの種類や発生といった音源情報を抽出することを主目的として設計されており，音源の方向や位置などの空間的情報は明示的には扱われていない．

2.2 空間拡張型音響言語モデル

音響言語モデルの空間拡張型として ELSA [1] が提案されている．ELSA は音源情報と空間情報を独立に扱う構成を採用しているため，複数の音源が存在する状況において，音源情報と空間情報の対応関係を表現することが困難である（Fig. 2a）．これに対し，Spatial-CLAP [5] は音源情報と空間情報が紐づいた埋め込み表現を学習することで，複数音源状況における両者の適切な対応関係を表現可能とする（Fig. 2b）．

2.3 自然言語生成における評価手法

文書要約や機械翻訳，キャプション生成といった自然言語生成タスクにおいては，参照文との一致度に基づ

*Generation and Evaluation of Audio Captioning Incorporating Spatial Information in Multi-Sound-Source Environments by Kentaro Seki (UToyo/Keio), Yuki Okamoto, Kouei Yamaoka (UTokyo), Yuki Saito (UTokyo/AIST), Shinnosuke Takamichi (Keio/UTokyo), Hiroshi Saruwatari (UTokyo).

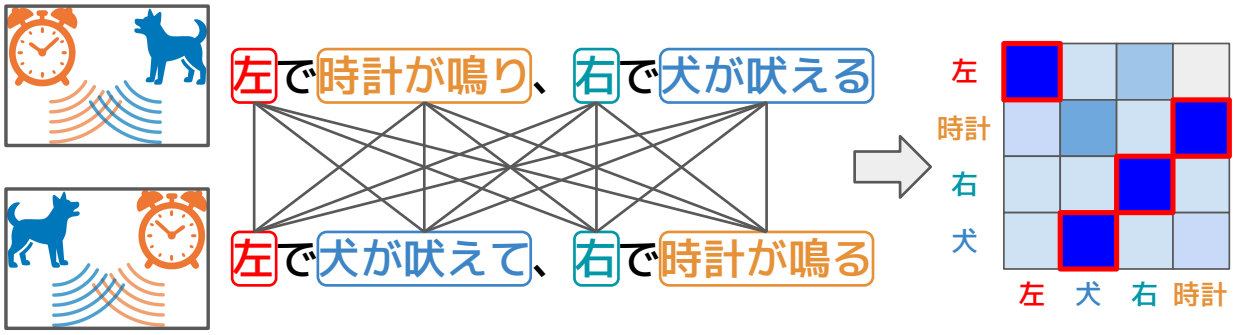


Fig. 3: 音源情報と空間情報の対応関係が異なるキャプションに対する，BERTScore [3] による評価の模式図．右側のマトリクスはキャプション間の類似度を表しており，濃い色ほど高い類似度を示す．両キャプションは構成要素が同一であるため高い類似度が得られるため，音源情報と空間情報の対応関係の違いが評価結果に反映されにくいと予想される．

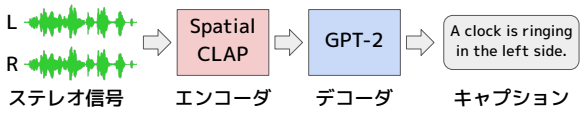


Fig. 4: 空間的音響キャプショニングのベースライン手法の全体像．

いて生成文の品質を評価する様々な手法が提案されている．古典的な評価手法は BLEU [6] のように一致度に基づく手法が主流であり，語順の違いやリフレーズといった表層的な語彙の差異に対する頑健性に欠ける．事前学習済みモデルによる埋め込み表現を用いる手法は，文の意味的類似度に基づく評価を可能とするため，表層的な違いに対してより頑健であることが期待される．代表的な手法として，SentenseBERT [2] や BERTScore [3] が挙げられる．しかし，これらの手法は文全体の意味的類似度を評価するものであり，細かな文意の差異を適切に評価することは困難であると考えられる（Fig. 3）．

3　空間的音響キャプショニング

3.1　ベースライン手法

本研究では，Spatial-CLAP [5] の音響エンコーダと GPT-2 [7] の言語デコーダを組み合わせた手法をベースラインとして用いる（Fig. 4）．音響エンコーダは入力のステレオ音響信号から音源情報と空間情報の統合された埋め込み表現を抽出し，線形層によって L 個のトークンに射影し，これを条件付けとして GPT-2 によりキャプションを生成する．

3.2　Spatial-CLAP による評価

SentenceBERT による評価と同様に，本研究では埋め込み表現間の類似度としてコサイン類似度を評価する．Spatial-CLAP の言語エンコーダによる参照文と生成文の n 次元埋め込みをそれぞれ $\boldsymbol{x}_{\mathrm{ref}}, \boldsymbol{x}_{\mathrm{gen}} \in \mathbb{R}^n$ とすると，コサイン類似度は次式で定義される：

$$\mathrm{cossim}(\boldsymbol{x}_{\mathrm{ref}}, \boldsymbol{x}_{\mathrm{gen}}) = \frac{\boldsymbol{x}_{\mathrm{ref}}^T \boldsymbol{x}_{\mathrm{gen}}}{\|\boldsymbol{x}_{\mathrm{ref}}\| \, \|\boldsymbol{x}_{\mathrm{gen}}\|}. \qquad (1)$$

Spatial-CLAP の言語エンコーダは空間対照学習により音源情報と空間情報の対応関係を学習している．そのため，参照文と生成文でこの対応関係が異なる場合には，類似度が低下すると期待される．

3.3　主成分分析を用いた評価

言語埋め込みの空間は異方的であり，一部の部分空間が大きな分散を持つことが知られている [8]．そのため，Spatial-CLAP の言語エンコーダを用いた評価においても，主成分分析を用いた部分空間への射影によってノイズを削減し，評価対象の要素を強調できると考えられる．

後述の手順によってキャプションを生成し，Spatial-CLAP の言語エンコーダによる埋め込みを主成分分析する．得られた平均ベクトルを $\boldsymbol{\mu}$ とし，共分散行列の固有ベクトルを大きい固有値に対応するものから順に $\boldsymbol{e}_1, ..., \boldsymbol{e}_n$ とする．ここで $\{\boldsymbol{e}_i\}_{i=1}^n$ は $\mathbb{R}^n$ の正規直交基底とする．埋め込みベクトル $\boldsymbol{x} \in \mathbb{R}^n$ に対し，k 次元部分空間への射影 $f_k(\boldsymbol{x})$ を次で定める：

$$f_k(\boldsymbol{x}) = \sum_{i=1}^{k} \left\{ (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{e}_i \right\} \boldsymbol{e}_i \qquad (2)$$

射影されたベクトル同士のコサイン類似度 $\mathrm{cossim}(f_k(\boldsymbol{x}_{\mathrm{ref}}), f_k(\boldsymbol{x}_{\mathrm{gen}}))$ により，参照文と生成文の類似度を評価する．

ここで，$f_k(\cdot)$ の構成は線形層の学習とみなせるため，主成分分析に用いるキャプションの作成には学習データを用いる．評価データを用いて主成分分析を行うと，データリークが生じる恐れがあるためである．

また，このキャプションの生成には，音源情報・空間情報の組み合わせ，音源情報のみ，空間情報のみ 3 つの場合を用意する．音源情報のみのキャプションで主成分分析を行うことで，音源情報に対応する部分空間での類似度を評価し，音源情報のみでの類似度評価が行われることが期待される．空間情報についても同様のことが期待されるため，これらの評価を組み合わせることで空間的音響キャプショニングの多角的な評価が可能と考えられる．

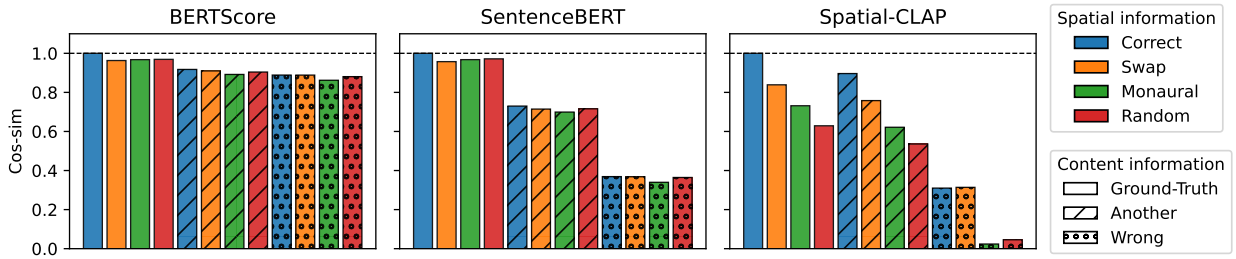


Fig. 5: 人工キャプションに対する評価結果．複数音源状況に固有の誤りである “Swap” について，従来手法では検出が困難であるが，Spatial-CLAP では “Correct” と区別して評価可能となる．

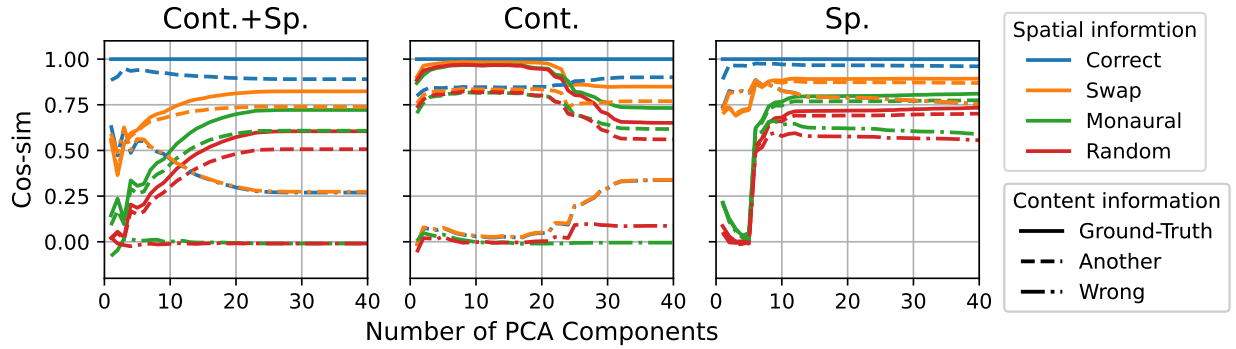


Fig. 6: 主成分分析に用いる次元数に対するコサイン類似度の変化．

4　実験 1：人工キャプションの評価

評価手法の適切さを検証するために，人工的に作成したキャプションに対する評価結果を比較する．

4.1　実験条件

AudioCaps 2.0 [9] のテストセットを用いる．1 つのクリップにつき 5 つのキャプションが付与されているため，本研究では事前に各クリップから 1 つのキャプションを選択する．各キャプションを 1 音源とみなし，2 音源に異なる空間情報を付与することで複数音源状況を模擬した空間的キャプションを作成し，参照文とする．

生成文では，音源情報に関して，参照文と同一文を用いる “Ground-Truth”，同一クリップに付与された異なるキャプションを用いる “Another”，および異なるクリップのキャプションを用いる “Wrong” の 3 通りを検討する．空間情報に関しては，正しい空間情報を付与する “Correct”，2 音源の空間情報を入れ替えた “Swap”，空間情報を付与しない “Monaural”，およびランダムな空間情報を付与する “Random” の 4 通りを検討する．

4.2　結果

4.2.1　従来手法との比較

BERTScore [3]，SentenceBERT [2] 埋め込みのコサイン類似度，および Spatial-CLAP [5] による言語埋め込みのコサイン類似度について，人工キャプションに対する評価結果を Fig. 5 に示す．BERTScore や SentenceBERT による評価では “Correct” と “Swap” の差が小さく，音源情報と空間情報の対応関係の精度を評価することが困難である．これに対し，Spatial-CLAP による評価では “Correct” と “Swap” の違いを明確に区別することが可能である．

4.2.2　主成分分析による評価

主成分分析の成分数 k に対する評価結果の違いを Fig. 6 に示す．ここで，“Cont.+Sp.“は音源情報と空間情報，“Cont.” は音源情報のみ，“Sp.” は空間情報のみを用いて主成分分析を行った場合を表す．低次元では手法間の評価値の差異が強調されており，特にその差異の生じ方が主成分分析に用いるキャプションの種類に依存することが分かる．そこで，適切な k を定めるため，比較したい手法群の間で評価値の差異が最大となる k を選択した．“Cont.+Sp.” では “Correct” と “Swap”，“Cont.” では (“Ground Truth,” “Another”) と “Wrong”，“Sp.” では (“Correct,” “Swap”) と (“Monaural,” “Random”) の間で差異を計算し，それぞれ $k = 2, 13, 4$ となった．

この k での手法間の比較結果を，Fig. 7 に示す．“Cont.+Sp.” では全体的なキャプションが正しい “Ground Truth + Correct” および “Another + Correct” のみが高い値を取り，他の手法は低く評価されており，総合的なキャプションの精度が低い手法に低いスコアを付与できていると言える．“Cont.” では，“Ground Truth” 群と “Another” 群が近い値を取りながら，“Wrong” 群は 0 に近い値を取っている．音源情報が同一の手法群の中では空間情報の違いによる差異は小さく，音源情報に注目した評価が実現していると言える．ただし，“Ground Truth” 群と “Another” 群の間には差異が認められる．これはキャプションの作成者によって記述の粗さが異なることに起因すると考えられ，音響キャプショニングという

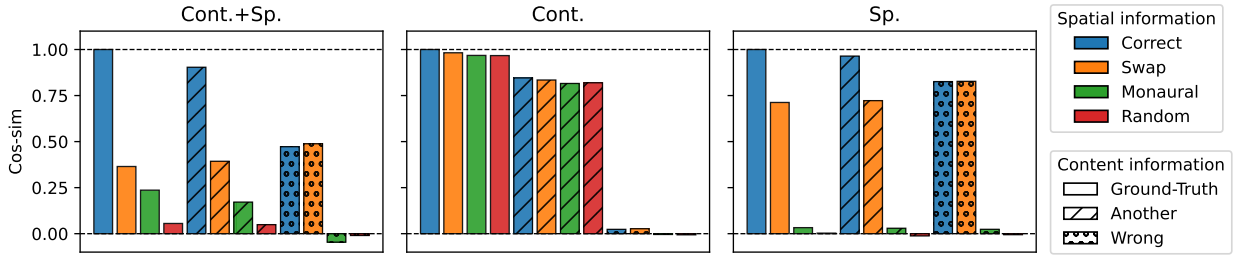


Fig. 7: 主成分分析を用いた評価による，人工的に作成したキャプションの評価結果．主成分分析に用いるキャプションによって，異なる性質の評価が実現することが分かる．

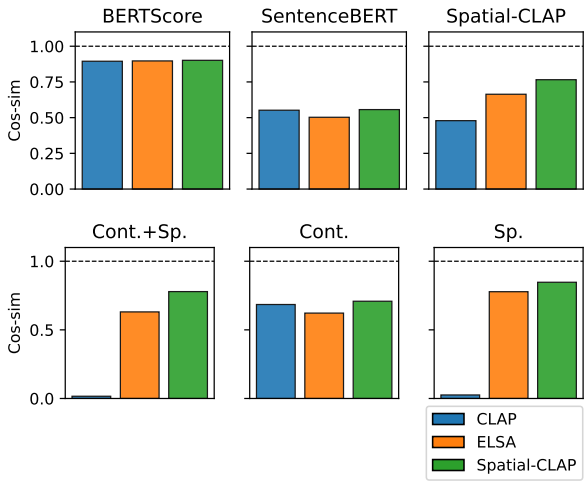


Fig. 8: 生成キャプションの評価結果．

タスクが持つ根源的なゆらぎを示している．“Sp.” では “Correct” 群と “Swap” 群が高い値を取りながら，“Monaural” 群および “Random” 群が 0 に近い値を取っている．この結果から空間情報に着目した評価が実現していると言えるが，“Correct” と “Swap” の間に差異が認められ，音源情報への依存が認められる結果となった．

5　実験 2：生成キャプションの評価

5.1　実験条件

Section 3.1 で説明したベースライン手法を，AudioCaps 2.0 [9] に空間情報を付与したデータセットによって学習・評価した．音響信号に対する空間情報の付与は，pyroomacoustics [10] による音響シミュレーションを用いた．音響エンコーダには Spatial-CLAP [5] を用い，そのパラメータは固定した．GPT-2 [7] の条件付けトークンの数は $L = 10$ とした．最適化には AdamW [11] を用い，学習率 1×10^{-5}，エポック数 50，バッチサイズ 128 とした．

比較手法として，音響エンコーダを従来の CLAP [12] および ELSA [1] に置き換えた手法を採用した．従来の CLAP はモノラル音響信号を入力とするため空間情報を扱うことができない．また，ELSA は音源情報と空間情報を独立に扱う構成を採用しているため，複数音源状況における両者の対応関係を明示的に表現することが困難である．

5.2　評価結果

生成キャプションに対する評価結果を Fig. 8 に示す．いずれの手法も音源情報を一定程度捉えているため，従来の意味的評価指標では手法間の差異は小さい．一方で，提案する評価手法では，空間情報の扱いに起因する差異が明確に現れている．特に “Sp.” では，モノラル音響信号を入力とする “CLAP” と，ステレオ音響信号を入力とする “ELSA” および “Spatial-CLAP” の間で大きな性能差が確認された．一方，“Cont.+Sp.” では “ELSA” が一定の性能を達成しており，音源情報と空間情報が完全には分離されずに表現されていることが示唆される．

6　まとめ

本研究では，複数音源状況下における空間的音響キャプショニングの課題を指摘し，Spatial-CLAP を用いたベースライン手法および評価手法の有効性を実験的に示した．今後の課題として，拡散性雑音を含むより複雑な音響環境への拡張が挙げられる．

7　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] Bhavika Devnani *et al.*, *Proc. NeurIPS*, vol. 37, pp. 33 505–33 537, 2024.
[2] Nils Reimers *et al.*, in *Proc. EMNLP-IJCNLP*, Nov. 2019, pp. 3982–3992.
[3] Tianyi Zhang *et al.*, in *Proc. ICLR*, pp. 1–43.
[4] Xinhao Mei *et al.*, *EURASIP journal on audio, speech, and music processing*, vol. 2022, no. 1, p. 26, 2022.
[5] 関 健太郎 *et al.*, in 日本音響学会秋季研究発表会, Sep. 2025.
[6] Kishore Papineni *et al.*, in *Proc. ACL*, 2002, pp. 311–318.
[7] Alec Radford *et al.*, *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
[8] Kawin Ethayarajh, in *Proc. EMNLP-IJCNLP*, 2019, pp. 55–65.
[9] Chris Dongjoo Kim *et al.*, in *Proc. NAACL-HLT*, 2019.
[10] Robin Scheibler *et al.*, in *Proc. ICASSP*. IEEE, 2018, pp. 351–355.
[11] Ilya Loshchilov *et al.*, *arXiv preprint arXiv:1711.05101*, 2017.
[12] Benjamin Elizalde *et al.*, in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.