社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

空間音とテキストの対照学習による音源情報と空間情報の分離表現学習

上治正太郎† 高道慎之介†,†† 山岡洸瑛††

† 慶應義塾大学
†† 東京大学

あらまし 本研究では，複数マイクで観測され，到来方向・距離・残響などの空間情報を含む多チャネル音響信号とテキストの対照学習において，音源情報と空間情報を分離して埋め込む二重埋め込み法を提案する．先行研究では音源情報と空間情報を単一の埋め込みを学習しており，どちらか一方の因子のみを扱うことは困難であった．我々はこの問題を，音源および空間に特化した二つの射影ヘッドを導入することによって解決する．これらのヘッドは，音源情報または空間情報のいずれか一方のみを共有する多チャネル音響信号とテキスト対を用いた対照学習によって獲得される．実験により提案法は各因子を分離して学習できることを示す．すなわち，音源埋め込みは音源に敏感かつ空間に鈍感であり，空間埋め込みはその逆になることを示す．

キーワード クロスモーダル埋め込み，空間音響，対照学習，分離表現学習

Shotaro UEJI†, Shinnosuke TAKAMICHI†,††, and Kouei YAMAOKA††

† Keio University
†† The University of Tokyo

1. ま え が き

複数マイクで観測され，到来方向・距離・残響などの空間情報を含む多チャネル音響信号（空間音）は，音源情報と空間情報という 2 つの属性を含んでいる．音源情報とは，音の種類，話者，出来事，音楽音・環境音といった内容的属性を指す [1]．一方，空間情報とは，到来方向，音源とマイクロホンの距離，残響など，位置や部屋の音響的な特徴に関する属性を指す [2]〜[4]．音とテキストの埋め込み表現を得るためには，対照学習 [5] が広く用いられている．この手法で得られた埋め込み表現は，音/テキスト検索や環境音分類など，さまざまな下流タスクでゼロショットに利用できる [6],[7]．しかし，従来の対照学習は単一チャネル音を対象としているため，音源情報の表現には優れる一方で，空間情報を扱うことは出来ない [7],[8]．

これに対して，ELSA（Embeddings for Language and Spatial Audio）[9] は，多チャネル音と，音源情報と空間情報を同時に記述するテキスト（例：“A woman speaking, originating to the right in a large echoey room”）との間で対照学習を行うことで，空間情報を考慮した音とテキストの埋め込み表現を学習する．この学習により，得られた埋め込み表現は音源情報と空間情報の両方を保持する．しかし，ELSA は音源情報と空間情報を単一の埋め込み表現にエンコードするため，これらの情報因子が絡み合い，それぞれを独立に扱うことが難しくなる．これに対して，図 1 に示すように，音とテキスト双方で，音源情報と空間情報を 2 つの別々の埋め込みとして学習することで，表現を分離して獲得する手法を提案する．具体的には，ELSA と同じエンコーダに対して，音源ヘッドと空間ヘッドの 2 つの射影ヘッドを追加し，弱教師の空間音とテキストのデータを用いた対照学習によってそれぞれを学習する．実験では，提案モデルは ELSA よ

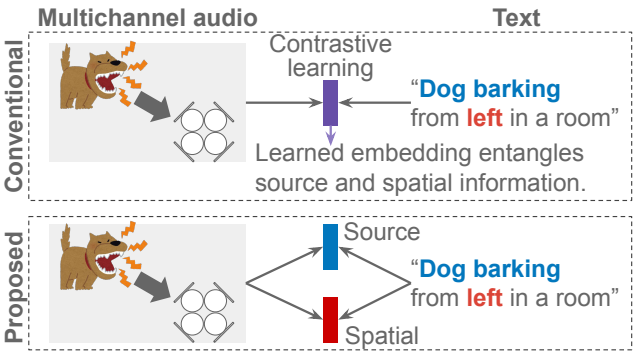

図 1: 従来の単一埋め込みモデル（上段）は音源情報と空間情報が混在するのに対し，提案する 2 分岐ヘッドモデル（下段）はそれらを音源埋め込みと空間埋め込みに分離する．

りも音源情報と空間情報に対する明確な分離性を示し，音源検索の性能を保持しつつ，off-task 検索（すなわち異なる因子をまたぐ検索）を強く抑制することが確認された．ソースコードは `https://github.com/takamichi-lab/disentangle`で公開している．

2. 関 連 研 究

2.1 非空間音とテキストの対照学習

多くの研究は，音をテキストと単一の埋め込み空間に写像させており，主に **音源** (*source*) 情報を対象としている．音とテキストの対照学習では，正例となる音とテキストの対を埋め込み空間で近づけ，負例のペアは遠ざけるように学習する．このようにして得られた埋め込みは，音/テキストに基づくゼロショット検索や，音分類などの下流タスクに利用できる [7], [8], [10]．これらは音源情報に対しては有効だが, 空間情報を扱うようには設計されていない．

2.2 空間音とテキスト：定位から表現へ

音イベント定位・検出（Sound Event Localization and Detection; SELD）は，音響イベント検出と音源定位を組み合わせたタスクである [2], [3]．近年の SELD 研究では自然言語による条件付けへと進展しているが，依然として焦点は定位であり，構造化された音 - テキストの埋め込みを学習する方向には向いていない [11]．ELSA は空間情報を考慮した音とテキストの埋め込みを学習するが，埋め込み内部では音源情報と空間情報が絡み合っている [9]．

2.3 音響分野における分離表現学習

因子の分離に関する研究では，説明因子を分離するためには帰納バイアスや何らかの教師信号が必要であることが指摘されている [12]．例えば，音声と環境の因子を分離する研究 [13] や，声における感情と話者を分離する研究 [14] がある．さらに，大域（global）と局所（local）の因子分離は，タスク間の転移性能を改善することが示されている [15]．これらの知見に基づき，本研究では空間音とテキストの対照学習における因子の分離を可能にし，各因子に対応した検索を実現するとともに，因子をまたぐ情報漏洩を抑制する．

3. 提 案 手 法

3.1 モデルの構造

本研究では，多チャネル音 $\boldsymbol{x}_{\text{aud}} \in \mathbb{R}^{M\times N}$ と，音源と空間の両方を記述するテキスト $\boldsymbol{x}_{\text{txt}}$ の間で対照学習を行う．ここで，M と N はそれぞれマイクロホン数とサンプル数を表す．各モダリティは，音源埋め込みと空間埋め込みの 2 種類を生成する．

音側では，音源エンコーダ $f^{(\text{src})}_{\text{aud}}(\cdot)$ として事前学習済み HTS-AT [16] を用い，空間エンコーダ $f^{(\text{spa})}_{\text{aud}}(\cdot)$ として事前学習済み spatial encoder [9] を用いる．テキスト側のエンコーダ $f^{(\text{enc})}_{\text{txt}}(\cdot)$ には，事前学習済み RoBERTa [17] を使用する．

両モダリティは，系列方向の平均化と多層パーセプトロン $f_{\text{aud}}(\cdot)$ および $f_{\text{txt}}(\cdot)$ を通して d 次元空間へ写像され，それぞれ ℓ_2 正規化された共有特徴 $\boldsymbol{y}_{\text{aud}}$ と $\boldsymbol{y}_{\text{txt}}$ を得る:

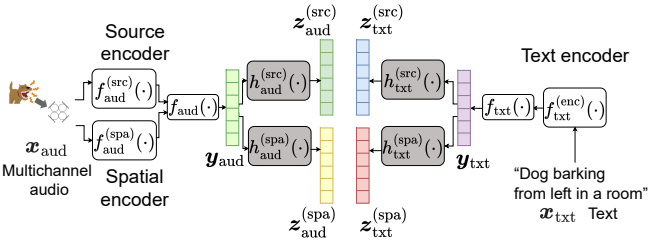

図 2: 提案モデルの構造．多チャネル音 $\boldsymbol{x}_{\text{aud}}$ とテキスト $\boldsymbol{x}_{\text{txt}}$ はそれぞれ音源と空間に共通する特徴量 $\boldsymbol{y}_{\text{aud}}$ と $\boldsymbol{y}_{\text{txt}}$ にエンコードされる．灰色のボックスは 2 分岐ヘッド構造 $h^{(*)}_{*}(\cdot)$ を示しており, 共有特徴量を音源情報を表現する埋め込み $(\boldsymbol{z}^{(\text{src})}_{\text{aud}}, \boldsymbol{z}^{(\text{src})}_{\text{txt}})$ と空間情報を表現する埋め込み $(\boldsymbol{z}^{(\text{spa})}_{\text{aud}}, \boldsymbol{z}^{(\text{spa})}_{\text{txt}})$ の計 4 種類へ写像する．

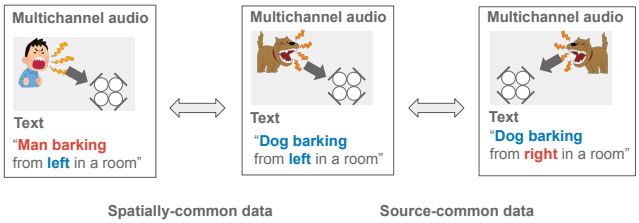

図 3: 2 つのペア (多チャネル音とテキストの対) 間の関係性：中央が参照ペア, 左は空間情報が一致していて音源が異なる空間共通 (spatially-common) データ, 右は音源が一致していて空間情報が異なる音源共通 (source-common) データ．

$$\boldsymbol{y}_{\text{aud}} = f_{\text{aud}}\left(\text{Concat}\left(f^{(\text{src})}_{\text{aud}}(\boldsymbol{x}_{\text{aud}}), f^{(\text{spa})}_{\text{aud}}(\boldsymbol{x}_{\text{aud}})\right)\right), \tag{1}$$

$$\boldsymbol{y}_{\text{txt}} = f_{\text{txt}}\left(f^{(\text{enc})}_{\text{txt}}(\boldsymbol{x}_{\text{txt}})\right), \tag{2}$$

ここで，Concat$(\cdot)$ は特徴軸での結合を表す． $f_{\text{aud}}(\cdot)$ と $f_{\text{txt}}(\cdot)$ の上に，音源情報と空間情報を明示的に分離するための 2 種類の射影ヘッドを付加する． 具体的には，2 つの音源ヘッド $h^{(\text{src})}_{\text{aud}}(\cdot)$，$h^{(\text{src})}_{\text{txt}}(\cdot)$ と 2 つの空間ヘッド $h^{(\text{spa})}_{\text{aud}}(\cdot)$，$h^{(\text{spa})}_{\text{txt}}(\cdot)$ から成る 4 つの多層パーセプトロンを用いる．各モダリティに対して，多層パーセプトロンは次のように出力を生成する．

$$\boldsymbol{z}^{(\text{src})}_{\text{aud}} = h^{(\text{src})}_{\text{aud}}(\boldsymbol{y}_{\text{aud}}), \qquad \boldsymbol{z}^{(\text{spa})}_{\text{aud}} = h^{(\text{spa})}_{\text{aud}}(\boldsymbol{y}_{\text{aud}}), \tag{3}$$

$$\boldsymbol{z}^{(\text{src})}_{\text{txt}} = h^{(\text{src})}_{\text{txt}}(\boldsymbol{y}_{\text{txt}}), \qquad \boldsymbol{z}^{(\text{spa})}_{\text{txt}} = h^{(\text{spa})}_{\text{txt}}(\boldsymbol{y}_{\text{txt}}). \tag{4}$$

各ヘッドの出力は ℓ_2 正規化された d 次元ベクトルであり，音源向け $(\boldsymbol{z}^{(\text{src})}_{\text{aud}}, \boldsymbol{z}^{(\text{src})}_{\text{txt}})$ と空間向け $(\boldsymbol{z}^{(\text{spa})}_{\text{aud}}, \boldsymbol{z}^{(\text{spa})}_{\text{txt}})$ の 4 種類の埋め込みが得られる．この 2 つのヘッド構造により，学習では目標とする因子（音源または空間）には感度を持たせつつ，非目標因子（空間または音源）に対しては不変になるよう促すことができる．

3.2 学習データ

本研究で用いる学習データは多チャネル音と記述文の対であり，各対は $(\boldsymbol{x}_{\text{aud}}, \boldsymbol{x}_{\text{txt}})$ である．テキストには，音源の記述と，方向・距離・部屋・残響などの空間的記述の両方が含まれる．

音源情報と空間情報を分離して学習するために，本研究では 2 つのペア間の関係を **空間共通** (*spatially-common*) と **音源共**

通 (*source-common*) の 2 種類で定義する．これらは図 3に示すとおりであり，いずれも室内音響シミュレーション（例: 鏡像法 [18]）によって合成可能である．

例えば，単一チャネル音とその記述文（例：犬の鳴き声と “dog barking”）の対を考える．学習データに含まれる対は，単一チャネル音に室内インパルス応答 (room impulse response; RIR) を畳み込み，さらに，記述文に空間に関する記述を追加することで生成される．

空間共通データは，2 つの異なる単一チャネル音と記述文対に同じ RIR と空間記述を適用することで得られる．逆に，同一の単一チャネル音と記述文対に異なる RIR と空間記述を適用することで，音源共通データが得られる．

3.3 学習の目的関数

教師あり対照学習 [19] によりモデルを学習する．$\{\ell_i^{(\mathrm{src})}\}_{i=1}^{B} \in \mathcal{Y}^{(\mathrm{src})}$ および $\{\ell_i^{(\mathrm{spa})}\}_{i=1}^{B} \in \mathcal{Y}^{(\mathrm{spa})}$ を，それぞれ音源と空間のラベルとする．ここで，$\ell_i^{(\mathrm{src})}$ は単一チャネル音と記述文対の識別子を，$\ell_i^{(\mathrm{spa})}$ は付与する RIR および空間記述に対応する識別子である．B はミニバッチサイズである．

音源ヘッドに対しては，第 i サンプルの正例インデックス集合を $\mathcal{P}^{(\mathrm{src})}(i) = \{j \in \{1,\dots,B\}\backslash\{i\} \mid \ell_j^{(\mathrm{src})} = \ell_i^{(\mathrm{src})}\}$. と定義する．空間ヘッドについては，$\mathcal{P}^{(\mathrm{spa})}(i)$ はラベルの “(src)” を “(spa)” に置換することで同様に記述される．ミニバッチは空間共通データと音源共通データの両方のサンプルで構成される．(i,j) 番目のサンプルが空間共通データであるならば，$j \notin \mathcal{P}^{(\mathrm{src})}(i)$ だが $j \in \mathcal{P}^{(\mathrm{spa})}(i)$ となる．一方，音源共通データでは $j \in \mathcal{P}^{(\mathrm{src})}(i)$ であり，かつ $j \notin \mathcal{P}^{(\mathrm{spa})}(i)$ となる．

音源ヘッドに対する対照損失は次のように記述される．

$$\mathcal{L}_{a\to t}^{(\mathrm{src})}(i) = \sum_{j\in\mathcal{P}^{(\mathrm{src})}(i)} \log \frac{\exp\left(\mathrm{s}\left(\boldsymbol{z}_{\mathrm{aud},i}^{(\mathrm{src})}, \boldsymbol{z}_{\mathrm{txt},j}^{(\mathrm{src})}\right)/\tau\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(\mathrm{s}\left(\boldsymbol{z}_{\mathrm{aud},i}^{(\mathrm{src})}, \boldsymbol{z}_{\mathrm{txt},k}^{(\mathrm{src})}\right)/\tau\right)}, \tag{5}$$

$$\mathcal{L}_{t\to a}^{(\mathrm{src})}(i) = \sum_{j\in\mathcal{P}^{(\mathrm{src})}(i)} \log \frac{\exp\left(\mathrm{s}\left(\boldsymbol{z}_{\mathrm{txt},i}^{(\mathrm{src})}, \boldsymbol{z}_{\mathrm{aud},j}^{(\mathrm{src})}\right)/\tau\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(\mathrm{s}\left(\boldsymbol{z}_{\mathrm{txt},i}^{(\mathrm{src})}, \boldsymbol{z}_{\mathrm{aud},k}^{(\mathrm{src})}\right)/\tau\right)}, \tag{6}$$

$$\mathcal{L}^{(\mathrm{src})} = -\frac{1}{|\mathcal{P}^{(\mathrm{src})}(i)|} \sum_{i=1}^{B} \left(\mathcal{L}_{a\to t}^{(\mathrm{src})}(i) + \mathcal{L}_{t\to a}^{(\mathrm{src})}(i)\right), \tag{7}$$

ここで，$\boldsymbol{z}_{\mathrm{aud},i}^{(\mathrm{src})}$ はミニバッチ内の i 番目の $\boldsymbol{z}_{\mathrm{aud}}^{(\mathrm{src})}$ を表す．同様に，$\boldsymbol{z}_{\mathrm{txt},j}^{(\mathrm{src})}, \boldsymbol{z}_{\mathrm{txt},i}^{(\mathrm{src})}, \boldsymbol{z}_{\mathrm{aud},j}^{(\mathrm{src})}$ もそれぞれ対応する成分を示す．さらに，$\mathcal{N}(i) = \{1,\dots,B\}\backslash\{i\}$ は分母に含まれるインデックス集合であり，$\mathrm{s}(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{u}^\top\boldsymbol{v}$ は埋め込みの類似度関数，$\tau > 0$ は学習可能な温度パラメータである．空間ヘッドの損失 $\mathcal{L}^{(\mathrm{spa})}$ は，上式の “(src)” を “(spa)” に置き換えることで同様に定義される．

ELSA に従い，本研究では音響物理量の回帰を目的関数に組み込む．具体的には，音源方向，マイクロホンと音源の距離，部屋の広さ，残響時間を，$\boldsymbol{z}_{\mathrm{aud}}^{(\mathrm{spa})}$ からそれぞれ別の多層パーセプトロン によって予測する．定式化については [9] を参照されたい．これらの物理量回帰の損失を $\mathcal{L}_{\mathrm{phy}}$ と記す．

最後に，全体の目的関数は 2 種類の対照学習項と物理量回帰項の加重和で定義し，重みは学習中にオンラインで適応させる [20].

$$\mathcal{L} = \lambda_{\mathrm{src}}\,\mathcal{L}^{(\mathrm{src})} + \lambda_{\mathrm{spa}}\,\mathcal{L}^{(\mathrm{spa})} + \lambda_{\mathrm{phy}}\,\mathcal{L}_{\mathrm{phy}}, \tag{8}$$

ここで，λ_{src}，λ_{spa}，λ_{phy} は各目的関数の重みを表す．

表 1: 物理メタデータから言語記述子へのマッピング. 最下段にはキャプションテンプレートが示されている。キャプションは，キャプションテンプレートから 1 つ選ばれ, {*orig*}, {*loc*} および {*room*} を埋めることによって生成される．ここで, *loc* は**距離**, **仰角**および**方向**を表す言語記述子をこの順序で連結し, *room* は**部屋の床面積**と**残響時間**を組み合わせたものである．{*orig*} は単一チャネル音（音源情報）のテキスト記述である．

空間的特徴	範囲	言語記述子
距離	< 1 m	*near/close/nearby*
	> 2 m	*far/distant*
方向 (方位角)	$[-35°, +35°]$	*front/in front*
	$[+55°, +125°]$	*right/to the right*
	$[-125°, -55°]$	*left/to the left*
	$< -145°, > 145°$	*back/behind*
仰角	$> 40°$	*up/above*
	$< -40°$	*down/below*
残響時間 (T_{30})	< 200 ms	*acoustically dampened/dry-sounding*
	> 1000 ms	*highly reverberant/echoey*
部屋の床面積	< 50 m^2	*small/tiny*
	$[50, 100]$ m^2	*mid-sized/medium*
	> 10 m^2	*large/spacious*
キャプションテンプレート	The sound: {orig} is coming from the {loc} of a {room} room.	
	You can hear {orig} from the {loc}, inside a {room} room.	
	The sound: {orig} originates in the {loc} of a {room} room.	

4. 実験的評価

4.1 実 験 設 定

データセット： ELSA に従い，本研究では AudioCaps [21] を拡張して空間音とテキストからなるデータセットを構築した．AudioCaps は単一チャネル音とそれに対応する英語キャプションから成る．各音クリップは，Pyroomacoustics [22] により生成した 4 チャネル RIR を用いて多チャネル音にした．畳み込んだ A-format 信号は，一次アンビソニック形式（FOA；$[W, Y, Z, X]$）へと射影し，サンプリング周波数 48 kHz，長さ 10 s の信号とした．表 1のように, 元のキャプションには，物理メタデータ（距離，方位角，仰角，部屋の広さ，残響時間 T_{30}）に基づいて生成した空間記述を挿入することで拡張を行った．

評価に際しては，距離（2），方位角（4），仰角（2），部屋の広さ（3），および T_{30}（2）の 5 因子からなるグリッドを構築し，これにより $2\times4\times2\times3\times2 = 96$ 種類の RIR を生成した．各ビンには 96 個の異なる単一チャネル音を対応付け，最終的に $96\times96 = 9216$ 個のテストデータを得た．

アーキテクチャとハイパーパラメータ： エンコーダは ELSA と同一であり，唯一のアーキテクチャ上の違いは，共有モダリティ埋め込みの上に 2 種類の射影ヘッドを追加した点である．4 種類の埋め込み $\boldsymbol{z}_{\mathrm{aud}}^{(\mathrm{spa})}, \boldsymbol{z}_{\mathrm{aud}}^{(\mathrm{src})}, \boldsymbol{z}_{\mathrm{txt}}^{(\mathrm{spa})}, \boldsymbol{z}_{\mathrm{txt}}^{(\mathrm{src})}$ はいずれも $\mathbb{R}^d$（$d = 512$）に属し，ELSA の音/テキスト埋め込み次元と一致している．モデルは 20 エポック学習し，バッチサイズは $B = 24$，ビュー数は $V = 3$ としたため，有効ビュー数は $24\times3 = 72$ と

表 2: モダリティごとの IIDRs. (注1)

モダリティ	モデル	埋め込み	$\mathrm{IIDR}^{(\mathrm{src})}$	$\mathrm{IIDR}^{(\mathrm{spa})}$
Audio	ELSA	-	1.6056	0.6228
	DISSE	Source	**5.5653**	0.1797
		Spatiality	0.2817	**3.5493**
Text	ELSA	-	1.7242	0.5800
	DISSE	Source	**15.5462**	0.0643
		Spatiality	0.1246	**8.0247**

なる．最適化には AdamW [23]（初期学習率 lr = 5×10^{-5}，重み減衰 wd = 0.01）を用い，線形ウォームアップ [24]，コサイン減衰スケジュール [25]，および自動混合精度 [26] を採用した．提案モデルである DISSE（Disentangled Source and Spatial Embeddings）では，RoBERTa $f_{\mathrm{txt}}^{(\mathrm{enc})}(\cdot)$，HTS-AT $f_{\mathrm{aud}}^{(\mathrm{src})}(\cdot)$，および空間エンコーダ $f_{\mathrm{aud}}^{(\mathrm{spa})}(\cdot)$（ELSA の論文 [9] の記述に基づきモデルを再実装し，同等のデータセットおよび学習設定で事前学習を行った）を凍結し，射影ヘッドと τ のみを最適化した．学習は NVIDIA A6000 GPU 1 枚で実施した．

4.2 音源と空間の分離性

4 種類の埋め込み（$\boldsymbol{z}_{\mathrm{aud}}^{(\mathrm{spa})}, \boldsymbol{z}_{\mathrm{aud}}^{(\mathrm{src})}, \boldsymbol{z}_{\mathrm{txt}}^{(\mathrm{spa})}, \boldsymbol{z}_{\mathrm{txt}}^{(\mathrm{src})}$）のそれぞれについて，目標因子には敏感でありつつ，非目標因子には鈍感となるかを評価する．この目的のため，本研究ではモダリティ内の埋め込み（音同士あるいはテキスト同士）に対して inter-/intra-class distance ratio（IIDR）[27] を算出した．IIDR は，平均クラス間距離を平均クラス内距離で割った値として定義される．ここでクラスとは，本論文では目標因子と非目標因子を指す．したがって，IIDR が大きいほど，目標因子のみに選択的に敏感であることを意味する．

目標因子が音源または空間であるとき，音側の IIDR は次のように定義される:

$$\mathrm{IIDR}_{\mathrm{aud}}^{(\mathrm{src})} = \frac{\mathbb{E}\left[d\left(\boldsymbol{x}_{\mathrm{aud}}^{(\mathrm{src},\mathrm{spa})}, \boldsymbol{x}_{\mathrm{aud}}^{(\mathrm{src}',\mathrm{spa})}\right)\right]}{\mathbb{E}\left[d\left(\boldsymbol{x}_{\mathrm{aud}}^{(\mathrm{src},\mathrm{spa})}, \boldsymbol{x}_{\mathrm{aud}}^{(\mathrm{src},\mathrm{spa}')}\right)\right]}, \tag{9}$$

$$\mathrm{IIDR}_{\mathrm{aud}}^{(\mathrm{spa})} = \frac{\mathbb{E}\left[d\left(\boldsymbol{x}_{\mathrm{aud}}^{(\mathrm{src},\mathrm{spa})}, \boldsymbol{x}_{\mathrm{aud}}^{(\mathrm{src},\mathrm{spa}')}\right)\right]}{\mathbb{E}\left[d\left(\boldsymbol{x}_{\mathrm{aud}}^{(\mathrm{src},\mathrm{spa})}, \boldsymbol{x}_{\mathrm{aud}}^{(\mathrm{src}',\mathrm{spa})}\right)\right]}, \tag{10}$$

ここで，$\mathbb{E}[\cdot]$ は該当する組にわたる期待値を表す．また，$\boldsymbol{x}_{\mathrm{aud}}^{(\mathrm{src},\mathrm{spa})}$ と $\boldsymbol{x}_{\mathrm{aud}}^{(\mathrm{src},\mathrm{spa}')}$ は音源共通データに属する音の組，$\boldsymbol{x}_{\mathrm{aud}}^{(\mathrm{src},\mathrm{spa})}$ と $\boldsymbol{x}_{\mathrm{aud}}^{(\mathrm{src}',\mathrm{spa})}$ は空間共通データに属する音の組を表す．$\mathrm{d}(\boldsymbol{a},\boldsymbol{b})$ は埋め込み $\boldsymbol{a}$ と $\boldsymbol{b}$ のコサイン距離を表す．

テキスト側の指標 $\mathrm{IIDR}_{\mathrm{txt}}^{(\mathrm{src})}$ および $\mathrm{IIDR}_{\mathrm{txt}}^{(\mathrm{spa})}$ は $\boldsymbol{x}_{\mathrm{aud}}$ を $\boldsymbol{x}_{\mathrm{txt}}$ に置き換えて同様に定義される．すべての IIDR はテストセット上で評価した．

表 2 は提案モデル DISSE の IIDR を示し，比較として単一埋め込みベースライン（ELSA）も併せて掲載している．ベースラインでは $\mathrm{IIDR}^{(\mathrm{spa})} < 1$ かつ $\mathrm{IIDR}^{(\mathrm{src})} > 1$ となり，音源情報に偏り，空間性に対する分離性が弱いことを示している．

(注1)：**太字** と 下線 はそれぞれ最大値と最小値を表す.

表 3: モダリティ内検索 (R@1 / MedR). (注2)

モデル	モダリティ	On-task (src)	On-task (spa)	Off-task (src)	Off-task (spa)
ELSA	aud→aud	0.919 / 1	0.080 / 19	0.919 / 1	0.080 / 19
	txt→txt	0.982 / 1	0.015 / 24	0.982 / 1	0.015 / 24
DISSE	aud→aud	**0.936** / 1	**0.703** / 1	0.274 / 13	0.064 / 97
	txt→txt	**0.994** / 1	**0.234** / 5	0.733 / 1	0.004 / 100

表 4: モダリティ間検索 (R@1 / MedR). (注2)

モデル	モダリティ	On-task (src)	On-task (spa)	Off-task (src)	Off-task (spa)	Both-task (src & spa)
ELSA	txt→aud	0.613 / 1	0.220 / 4	0.613 / 1	0.220 / 4	0.133 / 7
	aud→txt	0.643 / 1	**0.217** / 4	0.643 / 1	0.217 / 4	**0.142** / 7
DISSE	txt→aud	**0.631** / 1	**0.239** / 6	0.004 / 117	0.001 / 93	**0.135** / 8
	aud→txt	**0.647** / 1	0.144 / 17	0.008 / 95	0.001 / 88	0.115 / 10

一方で DISSE は，ベースラインに対して，目標因子側の値（音源埋め込みに対する $\mathrm{IIDR}^{(\mathrm{src})}$，空間埋め込みに対する $\mathrm{IIDR}^{(\mathrm{spa})}$）を増加させ，非目標因子側の値（空間埋め込みに対する $\mathrm{IIDR}^{(\mathrm{src})}$，音源埋め込みに対する $\mathrm{IIDR}^{(\mathrm{spa})}$）を減少させる．目標因子のスコアがベースラインより高く，非目標因子のスコアが低いことは，提案モデルが目標因子には高い感度を持ち，非目標因子には不変でいられることを示している．

さらにもう 1 つの観察として，テキストモダリティの IIDR は音モダリティよりも優れている．すなわち，目標因子の IIDR はより高く，非目標因子の IIDR はより低いという結果が得られている．この差を説明する仮説として，テキストエンコーダは，音源識別において音側のエンコーダよりも本質的に有利である可能性が挙げられる．例えば，「dog」と「cat」のテキスト表現は，それらに対応する音よりも一般に容易に分離できるためである．

4.3 音源/空間を考慮した検索

学習された埋め込み（音源埋め込みまたは空間埋め込み）が，目標因子（音源または空間）に対して高精度に検索でき，かつ非目標因子に対しては不正確であることを評価する．評価では，on-task 検索（例: 音源埋め込みを用いて音源情報が一致するデータを検索）と，off-task 検索（例: 空間埋め込みを用いて音源情報が一致するデータを検索）の 2 種類のタスクを実施した．この設定は “opposite-head retrieval” [28] に着想を得ている．

On-task および off-task 検索は，モダリティ内（intra-modal）とモダリティ間（cross-modal）の両方で評価した．評価項目は複数あり，4 種類のモダリティ方向（{aud, txt}→{aud, txt}）とターゲット（検索対象）となる 2 種類の因子（src, spa）から構成される．例えば，「aud→aud (src)」とは，音の埋め込みをクエリとし，音の埋め込みからなる検索候補の中から，ターゲットである音源情報が一致するデータを検索するタスクを意味する．

(注2)：太字は on-task（および both-task）において高いほど良い指標を示し，下線は off-task において低いほど良い指標を示す.

モダリティ内検索では，クエリと完全に一致する検索候補はマスクして除外した．ELSA では単一の埋め込みを使用した．一方，DISSE ではタスクおよび正解判定の基準となる因子に応じた埋め込み[注3] を利用した．

モダリティ間検索では，音源と空間の両方を一致させる both-task も用意した．このタスクでは，DISSE に対して連結した埋め込み Concat($z_{\text{txt}}^{(\text{src})}, z_{\text{txt}}^{(\text{spa})}$) および Concat($z_{\text{aud}}^{(\text{src})}, z_{\text{aud}}^{(\text{spa})}$) を使用した．

評価指標には Recall@K（R@K）と median rank（MedR）[29], [30] を用いた．On-task 検索で R@K が高いほど目標因子に対する検索性能が高く，off-task 検索で R@K が低いほど非目標因子からの情報漏洩が抑制されていることを意味する．

モダリティ内検索： 表 3 は，提案モデル DISSE のモダリティ内 R@K および MedR を示す．ELSA は音源に強く偏っており（“aud→aud (src)” および “txt→txt (src)” の R@K が高い），一方で空間性に関する性能は低い（“aud→aud (spa)” および “txt→txt (spa)” の R@K が低い）．これに対して，DISSE は on-task 検索において ELSA の値を上回り，off-task 検索において ELSA の値を下回った．特に on-task 検索における空間検索性能を大幅に改善した．これらの結果は，DISSE による分離が，目標因子に対する感度を高めつつ，非目標因子に対しては不変でいられることを示している．

さらに観察すると，on-task 検索においては “aud→aud (spa)” が “txt→txt (spa)” を大きく上回っている一方で，off-task 検索においては “txt→txt (src)” の性能が他のタスクに比べて低い．この違いを説明する 1 つの可能性として，事前学習されたエンコーダに関する非対称性が挙げられる．すなわち，音側には事前学習済みの空間エンコーダが含まれている一方で，テキスト側は空間的な基盤付けを持たない RoBERTa の事前学習に依存しているためである．

モダリティ間検索： 表 4 はモダリティ間検索の結果を示す．On-task における DISSE の性能は ELSA と同程度，もしくはわずかに上回る．重要なのは，DISSE が off-task 検索を強く抑制している点である．off-task における R@1 はほぼ 0 に近く，off-task 方向の情報漏洩が効果的に低減されていることを示す．さらに，both-task における性能はモデル間で同程度であり，分離された表現を用いても下流タスクの性能が損なわれていないことがわかる．

4.4 埋め込み表現の可視化

提案モデル DISSE の埋め込みが，目標因子には敏感でありつつ，非目標因子には不変であるかを定性的に評価する．4 種類の埋め込みを t-SNE [31] で可視化し，点を音源情報または空間情報ごとに色付けした．図 4 の 2×4 のグリッドにおいて，目標因子に対応する埋め込みは，(1,1), (2,2), (1,3), (2,4) の位置で密なクラスタを形成している一方，その他 (非目標因子に対応) では色が混在していることが確認できる．これは DISSE によって，目標因子を識別しつつ，非目標因子に対しては不変でいられることを示している．

(注3)：例えば off-task (src) では，空間埋め込みをクエリおよび検索候補の両方に用いて，音源情報を正解判定の基準とする．

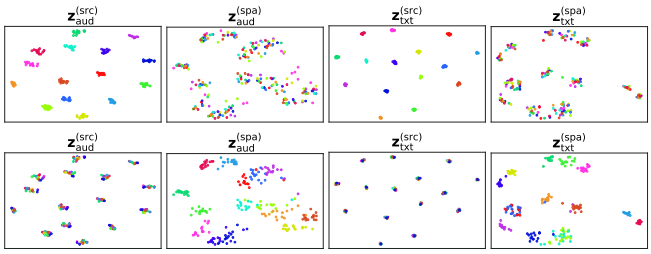

図 4: t-SNE による 4 つの埋め込みの可視化．上段は音源ごと，下段は空間ごとに色付けしている．

表 5: IIDR に関するアブレーション分析．[注1]

モダリティ	ELSA init.	Phys. loss	埋め込み	IIDR$^{(\text{src})}$	IIDR$^{(\text{spa})}$
Audio	✓	✓	音源	5.5653	0.1797
			空間	0.2817	3.5493
	✗	✓	音源	11.2535	0.0889
			空間	0.2647	3.7774
	✓	✗	音源	7.6247	0.1312
			空間	0.2376	4.2087
	✗	✗	音源	**20.2702**	0.0493
			空間	0.1803	**5.5454**
Text	✓	✓	音源	15.5462	0.0643
			空間	0.1246	**8.0247**
	✗	✓	音源	42.9277	0.0233
			空間	0.1338	7.4715
	✓	✗	音源	16.0056	0.0625
			空間	0.2088	4.7886
	✗	✗	音源	**67.6377**	0.0148
			空間	0.2196	4.5528

表 6: モダリティ間検索に関するアブレーション分析 (R@1 / MedR)．[注2]

ELSA init.	Phys. loss	モダリティ	On-task (src)	On-task (spa)	Off-task (src)	Off-task (spa)
✓	✓	txt→aud	0.631 / 1	0.239 / 6	0.004 / 117	0.000 / 93
		aud→txt	0.647 / 1	0.144 / 17	0.008 / 95	0.001 / 88
✗	✓	txt→aud	0.612 / 1	0.192 / 72	0.006 / 101	0.000 / 87
		aud→txt	0.640 / 1	**0.184** / 58	0.006 / 77	0.003 / 72
✓	✗	txt→aud	**0.636** / 1	**0.255** / 5	0.004 / 109	0.000 / 93
		aud→txt	**0.660** / 1	0.134 / 17	0.006 / 90	0.002 / 87
✗	✗	txt→aud	0.624 / 1	0.089 / 253	0.003 / 185	0.001 / 70
		aud→txt	0.624 / 1	0.088 / 188	0.003 / 122	0.007 / 61

4.5 アブレーション分析

本手法 DISSE をより深く理解するために ablation study を行った．検討した要素は以下のとおりである．

- **ELSA init.**：空間エンコーダ，音源エンコーダ，およびテキストエンコーダを ELSA の学習済み重みで初期化する設定．
- **Phys. loss**：学習の目的関数に $\mathcal{L}_{\text{phy}}$ を加える設定．

表 5 にしめす IIDR の結果を見ると，音側の埋め込みに関しては，ELSA init. と Phys. loss の両方を用いない手法が最良の性能を示した．一方で，テキスト側の空間埋め込みでは，両方の

要素を用いた手法がより良い性能を示した．全体として，IIDR の観点では両要素を省くことが有利な場合もあると言える．

表 6 に示す検索性能には以下の傾向が見られる．(i) On-task 検索では，ELSA init. を用いることでスコアが向上する．(ii) off-task 検索では総じてスコアが低いが，Phys. loss を用いた手法はさらにわずかに低い値を示し，望ましい傾向である．これらの結果は，ELSA init. が目標因子に対する感度を高める一方で，Phys. loss は因子間の情報漏洩を抑制する働きを持つことを示唆している．

5. ま と め

本研究では，音とテキストの表現において音源と空間の埋め込みを分離する DISSE を提案した．単一埋め込みのベースラインと比較して，DISSE は目標因子に対する感度を高めつつ，非目標因子に対しては不変性を保つことを実現し，因子ごとに特化した検索を可能にした．これにより，クロスモーダル表現学習において音源情報と空間情報を分離することの重要性を明らかにした．

謝辞：本研究は，JST 創発的研究支援事業 JPMJFR226V，JSPS 科研費 23K24895，23K28108，立石科学技術振興財団 研究助成 (S) の支援を受けて実施した．

文　献

[1] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *Proc. ICASSP*, pages 776–780, 2017.

[2] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2019.

[3] Archontis Politis, Annamaria Mesaros, Sharath Adavanne, Toni Heittola, and Tuomas Virtanen. Overview and evaluation of sound event localization and detection in DCASE 2019. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:684–698, 2021.

[4] Acoustics — measurement of room acoustic parameters — part 1: Performance spaces (iso 3382-1:2009). International Organization for Standardization, 2009. Defines reverberation metrics incl. T30.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, volume 119 of *PMLR*, pages 1597–1607, 2020.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, volume 139 of *PMLR*, pages 8748–8763, 2021.

[7] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *Proc. ICASSP*, 2023.

[8] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language–audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proc. ICASSP*, 2023.

[9] Bhavika Devnani, Skyler Seto, Zakaria Aldeneh, Alessandro Toso, Elena Menyaylenko, Barry-John Theobald, Jonathan Sheaffer, and Miguel Sarabia. Learning spatially-aware language and audio embeddings. In *Proc. NeurIPS*, 2024. Main Conference Track.

[10] Andrey Guzhov, Frederic Raue, Jörn Hees, and Andreas Dengel. AudioCLIP: Extending CLIP to image, text and audio. In *Proc. ICASSP*, pages 976–980, 2022.

[11] Jinzheng Zhao, Xinyuan Qian, Yong Xu, Haohe Liu, Yin Cao, Davide Berghi, and Wenwu Wang. Text-queried target sound event localization. In *Proc. EUSIPCO*, pages 261–265, Lyon, France, August 2024.

[12] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proc. ICML*, volume 97, pages 4114–4124. PMLR, June 2019.

[13] Ahmed Omran, Neil Zeghidour, Zalan Borsos, Félix de Chaumont Quitry, Malcolm Slaney, and Marco Tagliasacchi. Disentangling speech from surroundings with neural embeddings. In *Proc. ICASSP*, pages 1–5, 2023.

[14] Zongyang Du, Berrak Sisman, Kun Zhou, and Haizhou Li. Disentanglement of emotional style and speaker identity for expressive voice conversion. In *Proc. Interspeech*, pages 2603–2607, Incheon, South Korea, 2022.

[15] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Contrastive learning of global–local video representations. In *Proc. NeurIPS*, pages 7025–7040, 2021.

[16] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection. In *Proc. ICASSP*, 2022.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019.

[18] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979. Received June 6, 1978.

[19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Proc. NeurIPS*, 2020. Preprint available: arXiv:2004.11362.

[20] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proc. CVPR*, pages 7482–7491, 2018.

[21] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *NAACL-HLT*, pages 119–132. Association for Computational Linguistics, 2019.

[22] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *Proc. ICASSP*, pages 351–355. IEEE, 2018.

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. ICLR*, 2019.

[24] Priya Goyal, Piotr Dollár, Ross Girshick, and et al. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv:1706.02677*, 2017.

[25] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv:1608.03983*, 2016.

[26] Paulius Micikevicius, Sharan Narang, Jonas Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *Proc. ICML*, 2018. Preprint available: arXiv:1710.03740.

[27] Y. Zeng et al. scBiG for representation learning of single-cell gene embeddings. *NAR Genomics and Bioinformatics*, 2024.

[28] Hyeonggon Ryu, Seongyu Kim, Joon Son Chung, and Arda Senocak. Seeing speech and sound: Distinguishing and locating audios in visual scenes. In *Proc. CVPR*, 2025.

[29] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. CVPR*, 2015.

[30] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proc. BMVC*, 2018.

[31] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.