マルチモーダルな野鳥の検索を目的とした知識ベースの構築

寺本 優香 [1] 小島 諒介 [2]
[1] 同志社大学大学院文化情報学研究科　[2] 京都大学大学院 医学研究科
teramoto@mil.doshisha.ac.jp　kojima.ryosuke.8e@kyoto-u.ac.jp

概要

実フィールドでの野鳥観測は自然環境理解や環境モニタリングのために重要である．本研究では，野鳥の観測を補助するための野鳥検索システムの構築を目指し，キーワード検索に加え，知識グラフ，音響情報での検索を可能とするためのマルチモーダルな埋込みからなる知識ベースを構築する．これらの複数のモダリティを組み合わせることで，単一モダリティでは見つけることが困難な関係性を互いに補うことで，より柔軟な検索システムの構築が期待できる．本研究では，その第一歩として，個別のモダリティ間のから得られる埋め込みベクトルの違いを分析し可視化することで，複数のモダリティを鳥種検索に用いることの重要性を確認した．

1　はじめに

自然環境の理解や環境モニタリングのためには、周囲の動物や鳥の行動やコミュニケーションについて理解することが重要である。フィールドでの観測において野鳥の検索は，観測の補助を目的として古くから研究されており，有用なツールとして利用されている [1, 2]．野鳥の検索には主にキーワードをベースとした一般的な検索のほかに，野鳥の写真から検索を行う画像ベースのものと鳴き声から検査奥を行う音響ベースにした検索が存在している．特に，フィールドでの野鳥の検索においては，木などの障害物によって実際に写真等を取るのが難しい場合においても，実際に聞こえた鳴き声を録音し，その録音音声から検索をするといったアプリケーションが有用であるため，音響ベースでの検索や識別が広く研究されている．

これまでの既存の検索システムの多くは単一モダリティによる検索であり，複数のモダリティを用いた研究は限定的である．例えば，野鳥の画像識別においては，鳥種の検索をする際に，鳥種間の階層性や類似度を多面的に表現することの重要性が指摘されている [3]．そこで，画像識別のタスク以外の音声の検索においても分類系統やキーワード検索など複数のモダリティを統合した検索が重要であると期待できる．

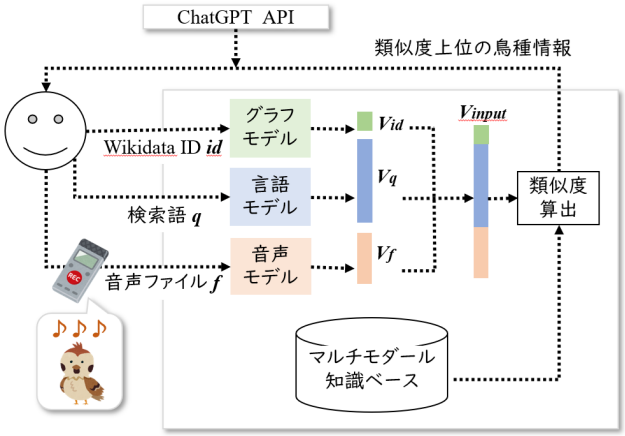


図 1　マルチモーダル検索システムの概要

これを踏まえ本研究では野鳥の知識グラフ・自然言語によるキーワード・音響の観点からの三種類の埋込みを統合し，鳥種ごとにマルチモーダルな埋込みを用いた検索システムを提案する．まず，一つ目のモダリティでは，知識グラフから取得したグラフ埋め込みを利用しており，カテゴリ間の関係が階層的な構造を持つ場合，より近縁関係にあるもの同士の類似度が高いことが期待される．Wu らは「シェパード」のラベルの判別ミスの例として「プードル」と「高層ビル」を挙げ，後者がより深刻であると述べている [4]．犬種と同様に鳥種も階層構造を保持しており，同様のアナロジーによってイエスズメとスズメのペアは，カラスとスズメのペアより近いと言える．従って系統樹による鳥種間の類似関係を議論することは重要である．二つ目のモダリティは，英語および日本語表記の鳥種の名称を BERT に入力し得られる単語埋込みを用いる．これは，近縁に属する鳥種の名称同士は部分的に同じ表記を含んでいたり，同じようなコンテキストで単語が出現する場合が多いということを利用している．例えばイ

エスズメとスズメは，同じスズメ科スズメ属に属する近縁種である．最後のモダリティは，鳥の鳴き声の音響データを用いて学習した wav2vec2 [1] によって得られた音声埋込みを用いる．鳥を含む生物全般の鳴き声の性質の要因として，体の大きさや構造，生息する環境が挙げられている [5]．音声埋込みを用いることで，こういった要因をもとにした類似関係の抽出が期待できる．

これらのことで，単一モダリティでは見つけることが困難な関係性を互いに補うことで，より柔軟な検索システムの構築が期待できる．本研究では，その第一歩として，個別のモダリティ間のから得られる埋め込みベクトルの違いを分析し可視化することで，複数のモダリティを鳥種検索に用いることの重要性を確認した

さらに本研究では複数のモダリティを組み合わせることの有効性を確かめるために，各モダリティごとに獲得した埋め込みベクトルの分析を行った．具体的には各モダリティごとに埋込みベクトルを作成し、鳥種間の類似度行列を作成した．この類似度を利用して，複数モダリティを複合した統合グラフを構築し，検索時に到達可能な情報のモダリティごとの違いについて考察を行った．この考察をもとにして，日英の単語埋込み・鳴き声の音声埋込み・知識グラフの埋込みを結合してマルチモーダルな埋込みを作成し，どのモダリティからでも検索可能な野鳥のマルチモーダル検索のための知識ベースを整備した．また，この知識ベースを用いて鳥種の検索システムのプロトタイプ実装を行った．

2　マルチモーダル検索システム

システムの全体像を図 1 に示す．本章ではシステム内部に格納された知識ベースの構築方法，およびシステム全体の解説を行う．

2.1　知識グラフ埋込み

本研究で用いる知識グラフは Wikidata[6] から取得したものを利用している．Wikidata は Wikipedia の情報を構造的に取り扱うことを目的として整備されており，各項目は Q から始まる識別子を持つとともに階層的に分類されている．ただし，Wikidata には野鳥以外の動物なども含まれているため，今回は「鳥」を表す項目である「bird(Q5113)」と，「bird(Q5113)」より下位概念に属する項目のみを抽出した．

ここで得られた知識グラフから node2vec[7] を用いて各ノードに関連付けられた埋め込みベクトルを取得し，本システムにおける検索に用いる．埋込みベクトルの次元数を 64 次元に設定し，node2vec のパラメータはランダムウォークのステップ数を 30，ウォーク総数を 200，計算時のスレッド数を 4 としている．さらに，モデル学習時のウィンドウサイズは 10，最小単語頻度は 1，パッチサイズは 4 とした．

2.2　単語埋込み

本システムでは英語と日本語の検索に対応するため，BERT を用いて，知識グラフに格納されている項目に含まれる全ての名称を単語埋込みに変換する．英語に対しては bert-base-uncased[2]，日本語に対しては bert-base-japanese-whole-word-masking[3] の学習済みモデルおよびトークナイザーを利用した．

2.3　音声埋込み

本研究において使用した鳥種の鳴き声の音声埋込み構築には，日本野鳥大鑑 [8] の付属音声データおよび，バードリサーチ[4]の公開データを使用している．これらの音声データを，wav2vec2.0[9] を用いて学習した[5](以降，単に wav2vec)．このモデルは，音声データを受け取り，音声埋込みのベクトルに変換している．ただし，時間フレームごとに得られたベクトルを平均したベクトルを用いている．また，同じ鳥種に複数の音声データが存在する場合，それらも平均化した．

2.4　鳥種検索用マルチモダール知識ベース

システム内部では，鳥種に関する知識グラフの情報が存在し，また各鳥種ごとに名称の単語埋込み（64 次元）・鳴き声の音声埋込み（768 次元）・知識グラフ埋込み（256 次元）を統合したマルチモーダルな埋込みが紐づけられている同じ鳥種に複数名称が存在する場合，名称数に対応したマルチモダールな埋込みが存在する．また，特定のモダリティが欠

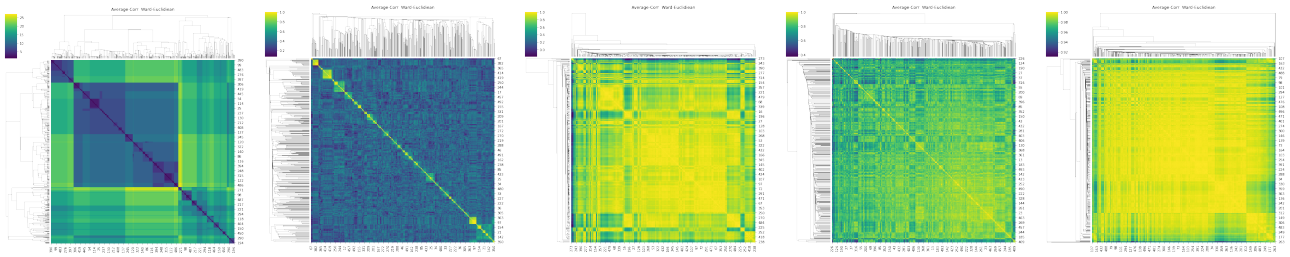
1）https://huggingface.co/kojima-r/wav2vec2-bird-jp-all

2）https://huggingface.co/bert-base-uncased

3）https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

4）http://www.bird-research.jp/

5）https://huggingface.co/kojima-r/wav2vec2-bird-jp-all



(a) 知識グラフを用いたノード間の距離行列 **(b)** 知識グラフを用いたグラフ埋込み間の類似度行列 **(c)** 英語名を用いた単語埋込み間の類似度行列 **(d)** 日本語を用いた単語埋込み間の類似度行列 **(e)** 鳴き声を用いた音声埋込み間の類似度行列

図 2　モダリティごとの鳥種間の距離行列および類似度行列

損している場合は，対応する埋込みの箇所をゼロベクトルで表現する．本稿ではこれらのデータセットをまとめて鳥種検索用マルチモダール知識ベースと呼ぶ．

2.5　マルチモーダル検索システム

図 1 中に示すように，本システムでは，1 種類または複数種類のモダリティの組合せを入力として受け取る．想定される入力は Wikidata の各項目ごとの識別子 id, 英語または日本語による検索キーワード q，鳥の鳴き声の音声データファイル f の 4 種類のうち．1 種類から全種類までの任意の組合せである．id が入力された場合，node2vec より V_id を得る．さらに入力として q が存在する場合は，q を BERT に入力し単語埋込み V_q を得る．ただし q が英語の場合は bert-base-uncased, 日本語の場合は bert-base-japanese-whole-word-masking を利用している．入力が音声ファイルであった場合，f に含まれる音声情報を wav2vec に入力し，音声埋込み V_f を得る．V_id，V_q，V_f の各埋込みの次元数はそれぞれ 64,768,256 であり，対応する入力が存在しなかった場合ゼロベクトルとなる．これらの埋込みを単純結合することで，入力を表現する埋込み V_input を作成している．V_input と知識ベースの全要素の類似度を算出し，最も類似度の高かった要素の ID を上位から検索結果として返す．さらに知識ベースを用い，検索された ID のノードにとって親ノード，子ノードとなる ID を取得し，検索された ID と入力との類似度とともに json 形式のデータにまとめ出力する．本システムは，json 形式の鳥種データを提示する GUI，ChatGPT API による自然言語での解説出力機能を備える．

3　分析結果

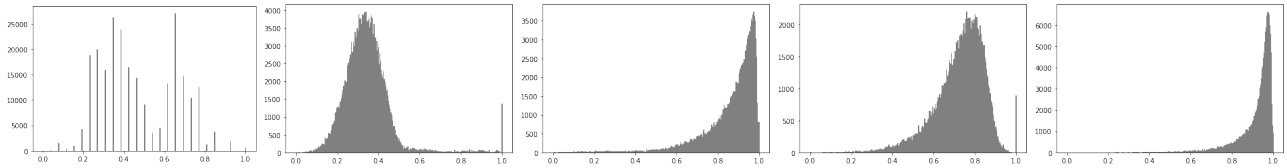
本章では，鳥種の概念を異なるモダリティから表現するこれらの特徴量を用い，鳥種間の類似度行列を作成した．それぞれの特性の差や利用方法について考察するため，これを正規化し，クラスタ分析を併せて行い，分析・可視化した．

3.1　埋込み同士の類似度行列と距離行列の計算

ここでは，各埋込みごとの性質および知識グラフの性質を分析するための処理を示す．図 1(a) は知識グラフにおける各ノード間の距離行列を正規化したものである．この距離は，二つのノード間の最短ホップ数で定義される．近いノード同士ほど距離が短くなり，同一のノード同士は距離は 0 になる．図 1(b) は知識グラフを node2vec により埋込み化したうえで，その埋込みの内積によって鳥種同士の類似度行列を算出した状態を示す．図 1(c) は英語の鳥種名称を BERT により埋込み化し，作成した類似度行列である．図 1(d) は同様の処理を日本語の鳥種名称に対し行っている．図 1(e) は wav2vec によって得られた音声埋込み同士の類似度行列を計算したものである．なお，図 1(b-e) に示す 4 種類の類似度行列については対角線上の同じ鳥種同士との類似度が 1 の値を取るよう正規化している．

これら 5 種類の行列に使用されている鳥種は比較のため．英語名称および日本語名称を持ち，音声データが存在し，さらに知識グラフ内で出現しているものに限定している．また，各図の上部および左部にあるのがクラスタデンドログラムであり，クラスタ分析を行った結果である．

知識グラフの距離行列から書く鳥種間の類似度を表現する行列を作成するため，1 から行列の各要素を引いた新たな行列を作成した．この距離に基づく類似度行列と，他の 4 種の類似度行列における値の



(a) 知識グラフの距離　**(b)** 知識グラフ埋込み　**(c)** 英語単語埋込み　**(d)** 日本語単語埋込み　**(e)** 音声埋込み

図 3　各モダリティごとの類似度行列の分布を示すヒストグラム

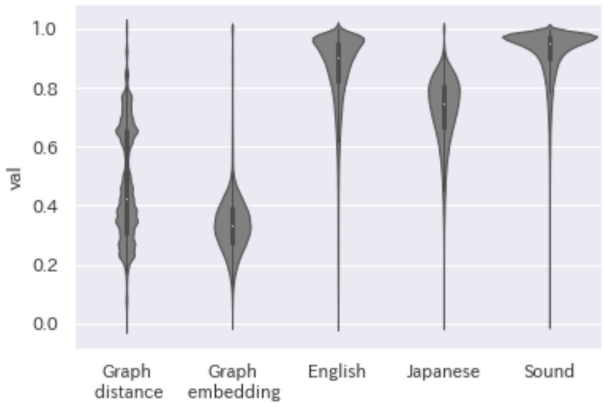


図 4　各モダリティごとの類似度行列の分布を示すヴァイオリンプロット

出現頻度を表すヒストグラムが図 3，バイオリンプロットが図 3 である．

3.2　各モダリティ間の関係

各モダリティごとに算出された鳥種間の関係性について分析を行うため，鳥種をノードとする埋込みグラフを構築した．具体的には，ノード間の類似度がある閾値以上のノード間にのみエッジを貼る処理を行った．前節で求めた知識グラフの距離に基づく類似度，知識グラフ埋込みに基づく類似度，日本語の言語埋込みに基づく類似度，英語の言語埋め込みに基づく類似度，音声埋込みに基づく類似度ごとに計 5 種類の埋込みグラフを構築した．著者の先行研究 [10] より，鳥種を表現するモダリティはその種類に関わらず，全体の 95%点よりも高い閾値を設定した場合，孤立したノードが増加することが判明している．

そこで，可視化のために閾値を 95%点とし，5 種類の埋込みグラフを統合したマルチモーダル埋込み統合グラフを作成した．このグラフの詳細については付録に示す．

4　考察

知識グラフの距離行列は，同じ値が多く出現する傾向があった．これは，グラフのもつ階層構造により，属する群によって距離がおおよそ決まるためであると考えられる．例えば，スズメ目スズメ科スズメ属に属する鳥種のノードから，タカ目タカ科の鳥種のノードへ移動するまでに通過するエッジの数は一定であり，各群に属する鳥種が増えるほどその距離が行列内に出現する．また知識グラフ埋込みの類似度行列は，単語埋込みや音声埋込みに近い値の分布を持つことが図より分かる．

英語の単語埋込みの類似度行列は，日本語に比べ互いに鳥種同士が近い傾向がある．これは，英語の鳥種名が複数の単語からなり，sparrow といった共通の単語を含みやすい傾向にあったためではないかと考えられる．さらに音声埋込みは他の類似度と比べ，最も類似度が高くなる傾向にあった．実際に音声検索を行った際，ほとんどの鳥種との類似度が高く，wav2vec の教師なしでの鳥種音声分類の困難さを反映している．

マルチモーダル埋込み統合グラフでは，1 種類のモダリティのみに着目した場合の埋込みグラフに出現するエッジ数が 9430 と最も多かった．ここから，特定のモダリティのみが検索可能な鳥種間の類似関係が多数存在することが明らかになった．従って，複数モダリティを併せて鳥種の検索に用いることは，様々な観点に基づいた類似の鳥種をユーザに提示するうえで重要であると言える．

5　おわりに

本研究では野鳥の鳴き声に注目したマルチモーダルな知識グラフの構築を行った．野鳥種間の類似関係を定量化するために英語・日本語の単語，鳴き声，Wikidata の知識グラフを活用した．またこれらを知識ベースとして整備し，マルチモーダル検索が可能な検索システムを構築した．今回作成した知識ベースを用いることで，鳥種間の類似度を定量的に扱うことが可能になる．このため，検索アプリケーションの提供以外に，鳥種の識別モデルにおける評価指標の一つとしての応用利用も期待される．

謝辞

本研究は JSPS 科研費 No.20H00475, 19KK0260 の助成を受けた。

参考文献

[1] A. Franzen and I.Y.H. Gu. Classification of bird species by using key song searching: a comparative study. In **SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat. No.03CH37483)**, Vol. 1, pp. 880–887 vol.1, 2003.

[2] 矢川雄一, 上野正俊, 田渕仁浩, 村岡洋一ほか. 鳥類図鑑 hyperbook における鳴き声検索. 全国大会講演論文集, データ処理, pp. 825–826, 1990.

[3] Shah Nawaz, Alessandro Calefati, Moreno Caraffini, Nicola Landro, and Ignazio Gallo. Are these birds similar: Learning branched networks for fine-grained representations. In **2019 International Conference on Image and Vision Computing New Zealand (IVCNZ)**, pp. 1–5. IEEE, 2019.

[4] Cinna Wu, Mark Tygert, and Yann LeCun. A hierarchical loss and its problems when classifying non-hierarchically. **Plos one**, Vol. 14, No. 12, p. e0226222, 2019.

[5] 阿部聖哉. 音声データによる野生生物調査の研究動向. 環境アセスメント学会誌, Vol. 18, No. 2, pp. 3–9, 2020.

[6] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. **Communications of the ACM**, Vol. 57, No. 10, pp. 78–85, 2014.

[7] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In **Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining**, pp. 855–864, 2016.

[8] 蒲谷鶴彦. 日本野鳥大鑑.

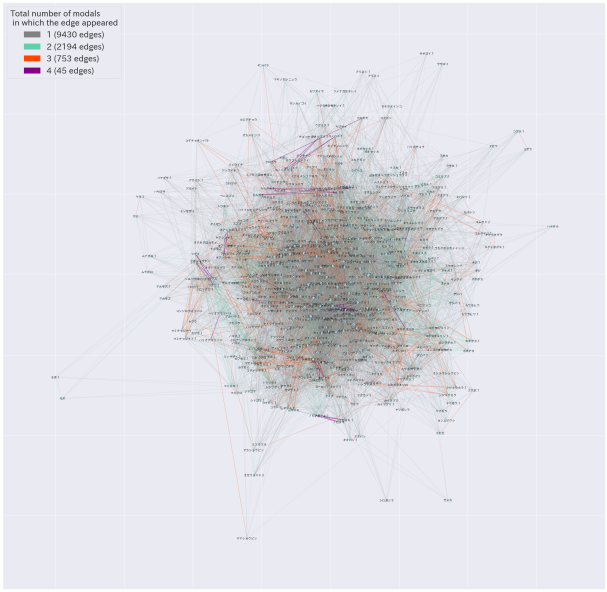
[9] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. **Advances in neural information processing systems**, Vol. 33, pp. 12449–12460, 2020.

[10] 寺本優香, 小島諒介. 複数マイクロホンアレイを用いた尤度分布統合による移動音源追跡. 日本ロボット学会, Vol. 2023, .

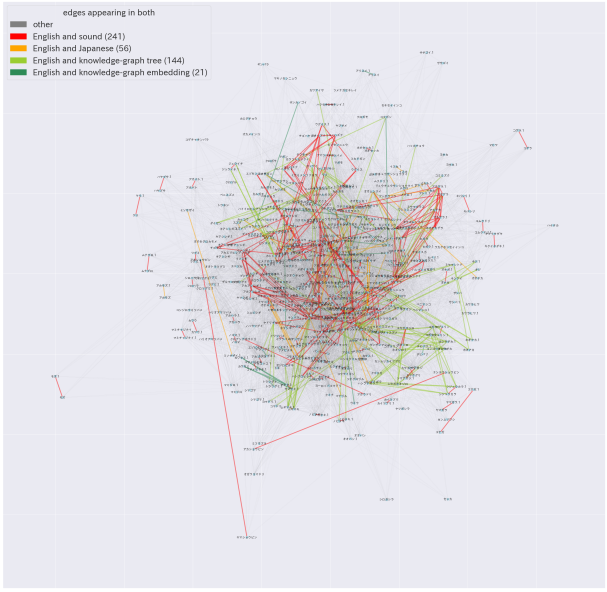
付録

マルチモーダル埋込み統合グラフは，5 種類の埋込みグラフに一度でも出現したエッジを全て保持している．A-a に表す全体像のエッジの色は，該当するエッジが知識グラフ・言語・音声のうち，何種類のモダリティに出現したかを表している．なお言語埋込みに関しては，日英いずれかまたは両方に出現したエッジを言語のモダリティにおいて張られたエッジとみなす．

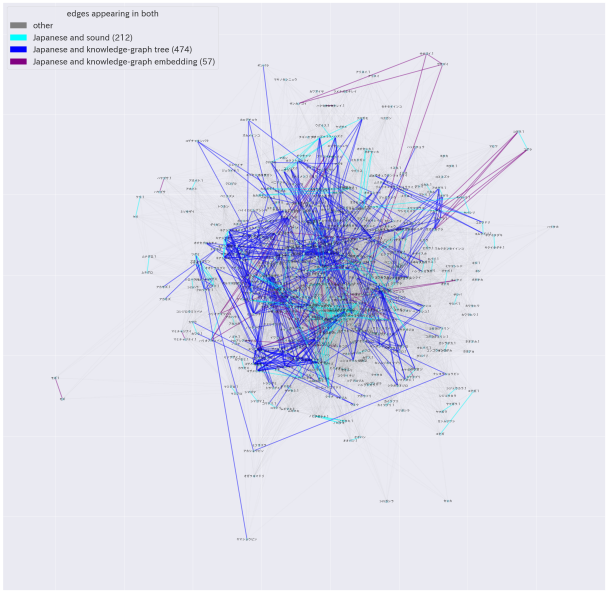
A-b,A-c,A-d は，2 種類のモダリティに共通して見られたエッジを可視化したものである．



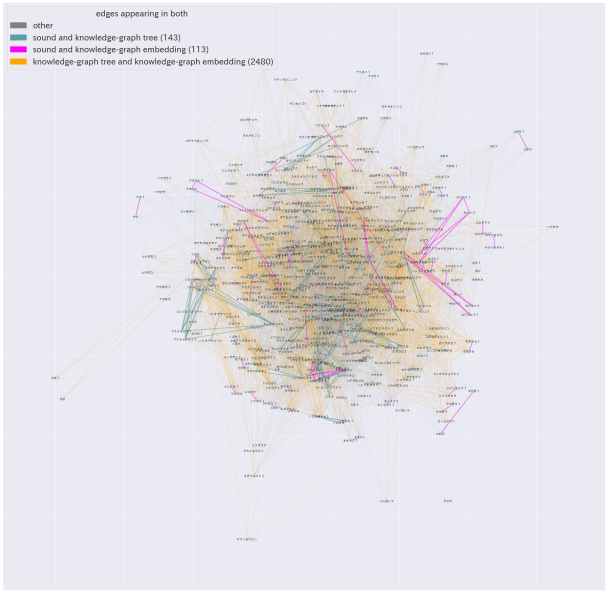
A-a マルチモーダル埋込みグラフ全体像



A-b 英語-音声（赤） 英語-日本語（橙） 英語-知識グラフ距離（黄緑） 英語-知識グラフ埋込み（緑）



A-c 日本語ー音声（水色） 日本語ー知識グラフ距離（青） 日本語ー知識グラフ埋込み（紫）



A-d 音声ー知識グラフ距離（緑） 音声ー知識グラフ埋込み（桃） 知識グラフ距離ー知識グラフ埋込み（金）