環境音に対する日本語自由記述文コーパスとベンチマーク分析

岡本 悠希 [1]　高道 慎之介 [2]　森松 亜依 [2]　渡邊 亞椰 [2]　井本 桂右 [3]　山下 洋一 [1]
[1] 立命館大学　[2] 東京大学　[3] 同志社大学
y-okamoto@ieee.org, shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp

概要

音の認識合成を大規模言語モデルと接続するためのデータセットの作成が急務である．本研究では，環境音データと，その内容を日本語で自由記述した文から成るオープンコーパスを構築する．本コーパスは環境音と英語自由記述文から成る既存コーパスの日本語訳であるため，英語と日本語を対比させた評価と分析が可能である．本論文は，コーパスの設計指針を述べるとともに，そのベンチマーク結果を分析する．

1　はじめに

大規模言語モデル（large language model, LLM）の隆盛を受け，テキスト以外のモーダルを LLM と接続して処理するためのデータセット構築が期待される．本論文では，環境音とテキストを相互変換できるタスクを扱い，そのための自由記述文–環境音データセットを構築する．この文は，発生音の内容物，順列，前後，前景背景などを自由記述したものであり，例えば

- **Text-to-audio**：自然言語から人工的に音を合成する，コンテンツ制作等に向けた技術 [1]–[3]
- **Audio captioning**：観測音の内容を自然言語で記述する，音情景理解に向けた技術 [4], [5]
- **Text-to-audio retrieval**：自然言語に基づき音データを検索する，データセット検索・作成に向けた技術 [6]
- **Text-audio model evaluation**：機械学習モデルが音の内容を理解しているかを評価 [7]

など多岐に渡る．これらを実現するデータセットの大半は英語 [8], [9] であり，それ以外には中国語のデータセット [10] が僅かに存在するのみである．しかしながら，既存研究で指摘されている音理解の不足 [7]，および，言語システムの違い（特に音象徴の言語間差異）について調査するためにも，様々な言語での学習および評価が必要である．

本研究では，日本語の自由記述文–環境音データセットを構築し，そのベンチマークを分析した結果を報告する．本データセットは，既存の英語データセット AudioCaps [8] を手動翻訳，および自動翻訳する形で作成する．本データセットは https://github.com/sarulab-speech/ml-audiocaps から入手可能である．ベンチマーク分析では，retrieval タスクを例にして，英語の既存研究で言及されている観点，及び日本語独特の音象徴に関する結果を報告する．

2　データセット作成

本研究で参考にするデータセット AudioCaps [8] は，audio captioning に向けて作成された英語データセットである．その音データは，希少イベントと楽音を除く多様な音イベントを含むように設計されている．英語の自由記述はクラウドソーシングにより作成されており，学習データ内の各音につき 1 つ，検証・評価データについては 5 つの自由記述文が付与されている．この自由記述文を日本語に翻訳することで，日本語のデータセットを作成する．なお，自動翻訳については学習データのみ，手動翻訳は全データを対象とする．

2.1　自動翻訳

自動翻訳のウェブサービスを利用して英語を日本語に翻訳する．品質を担保するため，翻訳した文に対し以下の処理を実施した．

- 翻訳元文の語句が残っている場合は，当該文をデータセットから削除する．
- 括弧などで注釈のついているものは，その括弧フレーズを削除（例えば，“トイレの流水（トイレの水を流すこと）→ トイレの流水”）する．ただし，注釈ではなく語の言い換えと見做される場合には，変更しない．
- 文が日本語として破綻している場合は，当該文を

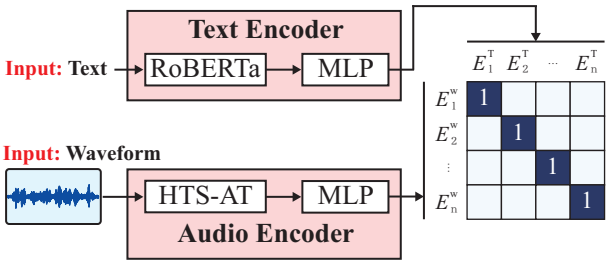

図 1　CLAP に基づく text-audio モデル構築の概要

データセットから削除する．

2.2　手動翻訳

以下の条件に基づき翻訳した．

- 文体は体言止めを基本とし，体言止めの記述が難しい場合には用言止めとする．
- 意味の似た英語名詞は、単一の日本語名詞に置き換える．例えば，“automobile, car” は “車” に翻訳する．
- 並列表現は句読点で結ぶ．例えば，“‘PhraseA and PhraseB” は “句 A，句 B” に翻訳する．
- 前景背景や順序を表す接続詞（例えば，while, after which）は，その役割を明示する表記（例えば，“同時に，その後で”）に翻訳する．
- 直訳としてオノマトペを伴うことが妥当である場合は，オノマトペを伴う．例えば，baaing, cackling, cawing はそれぞれ，“メーと鳴く”，“クワックワッと鳴く”，“カーカー鳴く” に翻訳する．これらのオノマトペは，対応する環境音の音象徴ではなく，あくまで元文の直訳であることに注意する．
- 元文が英語として破綻している，あるいは内容が不明確である場合には，当該元文を翻訳しない．

3　ベンチマーク

作成したデータセットのベンチマークとして，text-to-audio retrieval，audio-to-text retrieval タスク[1)]を実施した．まず，環境音と自由記述文の対応を獲得する深層学習モデル (以下，text-audio モデル）を学習した．モデルには，対照学習によって音とテキストを同じベクトル空間上に埋め込むように学習させる CLAP [11] を使用した．図 1 にモデルの概要を示す．Audio encoder と text encoder には，それぞれ hierarchical token semantic audio transformer (HTS-AT) [12] と RoBERTa [13] 使用した．なお，HTS-AT は CLAP が公式に提供している事前学習済モデル[2)]を使用した．また RoBERTa は，日本語，英語でそれぞれ学習された japanese-roberta-base[3)] 並びに roberta-base[4)] を使用した．学習には，日英それぞれの自由記述文と環境音のペアデータを用いて，言語ごとにモデルを構築した．

1）　Text-to-audio retrieval タスクと逆に，環境音から，対応する自由記述文を検索するタスク

表 1　自由記述文の元文と翻訳した日本語文の例

Language	Caption
English	• A dog barks twice and then whimpers
Japanese	• 犬の吠え声。その後に、クンクン鳴く音
Japanese (auto)	• 犬が 2 回吠えてから泣き声をあげます。
English	• A man talks while different birds tweet
Japanese	• 男性が話している声、同時に異なる鳥たちがチッチッと鳴く声
Japanese (auto)	• 男性が話す間に、異なる鳥たちがさえずります。

Retrieval の際は，学習させた audio encoder，text encoder をそれぞれ用いて入力データに対する埋め込みと検索対象データの埋め込み集合間のコサイン類似度を計算する．コサイン類似度が高いほど検索結果が上位であることを意味し，これらの結果をもとに環境音と自由記述文それぞれを入力とした場合にそれぞれ適切なデータを検索可能であるか評価する．

Retrieval の評価指標には，retrieval タスクにおいて使用される mean average precision at top 10 (mAP@10) 並びに，Recall at k (Recall@k）を使用した．Recall@k は，上位 k 番目までの retrieval 結果を全てのクエリ（text-to-audio retrieval の場合は text がクエリとなる）に対して平均した再現性スコアである．

4　結果

4.1　データセット

自動翻訳には OpenAI ChatGPT API (“gpt-3.5-turbo”) を用いた．アクセス時期は 2023 年 5 月である．“gpt-4” 及び “gpt-4-turbo” を使用しなかったのは，アクセス当時に当該 API が公開されていなかったためである．プロンプトには，翻訳を指示する文を含めたが，翻訳例や文ドメインに関する指示文は含めなかった．自動翻訳文の後処理，および手動翻訳は，一人の日本語母語話者が実施した．

2）　https://github.com/LAION-AI/CLAP
3）　https://huggingface.co/rinna/japanese-roberta-base
4）　https://huggingface.co/roberta-base

表 2　audio-to-text と text-to-audio retrieval の評価結果

Language	A → T				T → A			
	R@1	R@5	R@10	mAP@10	R@1	R@5	R@10	mAP@10
English	0.283	0.658	0.791	0.441	0.300	0.658	0.796	0.450
Japanese	0.234	0.592	0.751	0.386	0.223	0.571	0.744	0.375
Japanese (Auto)	0.237	0.555	0.689	0.368	0.229	0.544	0.701	0.364

表 3　並列表現を表す語句の先後を入れ替えた際の audio-to-text と text-to-audio retrieval の評価結果

Coordinate structure	Language	A → T				T → A			
		R@1	R@5	R@10	mAP@10	R@1	R@5	R@10	mAP@10
Original	Japanese	0.317	0.714	0.819	0.488	0.327	0.737	0.825	0.491
	Japanese (Auto)	0.333	0.683	0.794	0.474	0.325	0.665	0.807	0.471
Swap	Japanese	0.291	0.675	0.809	0.452	0.291	0.688	0.784	0.448
	Japanese (Auto)	0.319	0.675	0.796	0.465	0.296	0.670	0.778	0.446

検証，評価データセットはそれぞれ 2,475 文，4,875 文であり[5]，AudioCaps と同数である．一方，AudioCaps の学習データセットは 49,838 文であるのに対し，自動翻訳文と手動翻訳文の数はそれぞれ 49,668 文，48,947 文である．これは，2 章に示したように，不適データを削除したためである．

なお，AudioCaps の音データも我々のリポジトリで配布している．AudioCaps のリポジトリは音データを YouTube link として配布しており，動画の削除などによるデータの欠損が発生していた．これ以上の欠損を予防するため，我々が音データとして配布している．YouTube link からのダウンロード時期は 2023 年 11 月である．各音データの時間長は 10 秒であり，フォーマットは MP3 形式，48 kHz サンプリング，256 kbps とした．

4.2　ベンチマーク

3 章で構築したモデルを利用して，audio-to-text retrieval (A → T) 並びに text-to-audio retrieval (T → A) タスクを実施した．評価には，各言語によって学習されたモデルに対して，同一言語の評価データを使用した．なお，AudioCaps は 1 つの音につき 5 文の自由記述文が付与されているため，その中から 1 文をランダムに選択して，評価時の正解データとして扱った．以降，text-audio モデルは固定して，評価データのみ様々変化させて評価を実施する．

4.2.1　全体的な性能評価

表 2 に結果を示す．英語 (English) によって学習されたモデルと日本語によって学習されたモデルを比較すると，英語のほうが各指標において高いスコアを示した．Text Encoder に使用した RoBERTa は，英語モデルで約 160GB のテキストデータ，日本語モデルで約 75GB のテキストデータで学習されている．このように学習に使用されたデータ量が異なるため，それに起因して性能に差が見られたのではないかと考えられる．また，手動翻訳 (Japanese) と自動翻訳 (Japanese (Auto)) それぞれで学習したモデルの retrieval 結果を比較すると，手動翻訳が自動翻訳にやや勝るものの，その差は数% 程度である．そのため，自動翻訳されたデータセットも学習に活用できる可能性がある．

4.2.2　並列表現に関する評価

テキスト中の並列表現を表す語句を text-audio モデルがどの程度捕捉できているかについて検証した．評価には，4.2.1 節で使用した評価データセットの中から，並列表現（例えば “並行して”）を含む自由記述文のみを抽出した評価セット (Original) と，text 中の並列表現で区切られた句をランダムに入れ替えた自由記述文の評価セット (Swap) を用意した[6]．

表 3 に結果を示す．結果より，並列表現で区切られた句を入れ替えると，retrieval の性能がわずかに低下することが確認された．しかしながら，各指標において大きなスコア差は確認できなかった．これらの結果より，文中の並列表現で区切られた句を入れ替えても，retrieval の性能には大きな影響を与えないことが確認された．よって，日本語の自由記述文で学習した text-audio モデルにおいて，文中の並列表現を捕捉可能であることが明らかとなった．

5)　1 つの環境音につき 5 文が付与されているため，環境音データの数はこれらを値を 5 で割った数であることに注意する．公開リポジトリでは，この環境音データの数を表記している．

6)　例えば，“車両の走行音、男性と女性の話し笑う声” のような文の並列表現で区切られた句を入れ替えて，“男性と女性の話し笑う声、車両の走行音” という文を作成した

表 4　順序を表す語句の先後の句を入れ替えた際の audio-to-text と text-to-audio retrieval の評価結果

Order structure	Language	A → T				T → A			
		R@1	R@5	R@10	mAP@10	R@1	R@5	R@10	mAP@10
Original	Japanese	0.399	0.729	0.854	0.534	0.357	0.708	0.848	0.509
	Japanese (Auto)	0.384	0.717	0.819	0.522	0.342	0.702	0.807	0.485
Swapped	Japanese	0.330	0.688	0.819	0.486	0.286	0.664	0.792	0.442
	Japanese (Auto)	0.366	0.682	0.819	0.504	0.316	0.679	0.813	0.475

表 5　text 中からオノマトペを削除したデータに対する audio-to-text と text-to-audio retrieval の評価結果

Onomatopoeia	Language	A → T				T → A			
		R@1	R@5	R@10	mAP@10	R@1	R@5	R@10	mAP@10
Original	Japanese	0.234	0.592	0.751	0.386	0.223	0.571	0.744	0.375
	Japanese (Auto)	0.237	0.555	0.689	0.368	0.229	0.544	0.701	0.364
Eliminated	Japanese	0.240	0.563	0.729	0.379	0.206	0.548	0.719	0.356
	Japanese (Auto)	0.233	0.526	0.665	0.355	0.213	0.523	0.673	0.342

4.2.3　順序関係に関する評価

英語の既存研究 [7] おいて，文中の “before” や “after” といった環境音発生の順序を表す語句の先後の句を入れ替えても，retrieval タスクにおいて性能が低下しないことが報告されている．このことは，text-audio モデルが順序関係を学習できていないことを表している．そこで，日本語の自由記述文に対しても同様の実験を実施して，順序関係を補足可能であるか評価した．評価には，4.2.1 節で使用した評価データセットの中から，順序を表す語句（例えば “その後で”，“それに続いて”）を含む自由記述文のみを抽出した評価セット (Original)，文中の順序を表す語句の先後の句を入れ替えた自由記述文の評価セット (Swapped) を用意した[7)]．

表 4 に結果を示す．結果より，手動翻訳の自由記述文で学習させたモデルは，順序を表す語句の先後の句を入れ替えると mAP@10 が A → T で約 0.048，T → A で約 0.067 の性能低下が確認された．また，自動翻訳の自由記述文によって学習させたモデルは，mAP@10 が A → T で約 0.018，T → A で約 0.01 の性能低下であり，手動翻訳のモデルと比べると性能低下の幅が小さいことが確認された．表 1 に示す手動，自動翻訳による自由記述文を確認すると，手動翻訳の文は「句 A．その後に，句 B」のように順序が明示的であるのに対して，自動翻訳の方は，手動翻訳と比べて先後関係が取りにづらい文になっていることが確認できる．そのことに起因して，自動翻訳のモデルは手動翻訳の自由記述文で学習させたモデルよりも順序関係を捕捉しづらいモデルになったと推察できる．しかしながら，両者とも大幅な性能低下は確認されなかった．よって，英語の既存研究で示された結果と同様，日本語の自由記述文を用いた text-audio モデルの学習においても，text 中の順序関係を捕捉できないことが明らかとなった．

4.2.4　オノマトペに関する評価

日本語独特の音象徴としてオノマトペが挙げられる．今回翻訳した自由記述文はオノマトペを含むものも多く存在するため，text-audio モデルがこれらのオノマトペを retrieval の手がかりとして活用しているかどうかについて検証した．評価には，4.2.1 節で使用した評価データセットを利用して，文中からオノマトペを削除した自由記述文の評価セット (Eliminated) を用意した．

表 5 に結果を示す．結果より，文中からオノマトペを削除すると，各指標においてスコアがわずかに下がる傾向が確認された．よって，一部のデータにおいては，text-audio モデルがオノマトペを手がかりに retrieval を行ってると考えられる．ただし，2.2 節に述べたように，本データセットのオノマトペは環境音を受聴して付与したものではなく，英語の直訳である．そのため本実験の結果からでは，モデルが環境音の音象徴を直接利用したかを議論できないことに注意する[8)]．今後，どのような文からオノマトペを削除すると retrieval の際のスコアに変化が現れるかを詳細に分析する必要がある．

5　まとめ

本研究では，環境音を自由記述した日本語文コーパスの構築とベンチマーク結果を述べた．

7)　例えば，“ノックの音．それに続いて、機械のこぎりの音” のような文の先後の句を入れ替えて，“機械のこぎりの音．それに続いて、ノックの音” という文を作成した

8)　例えば，カエルの “クワックワッ” という鳴き声と “クワックワッ” のオノマトペが，本モデルの埋め込み空間で対応するか否かは，本実験結果からは議論できない

謝辞： 本研究は，ムーンショット JPMJPS2011，JST 創発的研究支援事業 JP23KJ0828，科研費 21H05054，21H04900，22H0363，23H03418，23K16908 の支援を受け実施した.

References

[1]D. Yang, J. Yu, H. Wang, *et al.*, "Diffsound: Discrete diffusion model for text-to-sound generation," *arXiv preprint arXiv:2207.09983*, 2022.

[2]F. Kreuk and G. Synnaeve and A. Polyak and U. Singer and A. Défossez and J. Copet and D. Parikh and Y. Taigman and Y. Adi, "AudioGen: Textually guided audio generation," in *Proc. International Conference on Learning Representation (ICLR)*, 2023 (to appear).

[3]H. Liu, Z. Chen, Y. Yuan, *et al.*, "AudioLDM: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[4]R. Mahfuz, Y. Guo, and E. Visser, "Improving audio captioning using semantic similarity metrics," in *ICASSP*, 2023, pp. 1–5.

[5]M. Kim, K. Sung-Bin, and T.-H. Oh, "Prefix tuning for automated audio captioning," in *ICASSP*, 2023, pp. 1–5.

[6]B. Elizalde, S. Deshmukh, M. A. Ismail, *et al.*, "CLAP: Learning audio concepts from natural language supervision," *arXiv:2206.04769*, 2022.

[7]H.-H. Wu, O. Nieto, J. P. Bello, *et al.*, "Audio-text models do not yet leverage natural language," in *ICASSP*, 2023, pp. 1–5.

[8]C. D. Kim, B. Kim, H. Lee, *et al.*, "AudioCaps: Generating captions for audios in the wild," in *NAACL-HLT*, 2019.

[9]K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *ICASSP*, 2020, pp. 736–740.

[10]X. Xu, H. Dinkel, M. Wu, *et al.*, "Audio caption in a car setting with a sentence-level loss," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021, pp. 1–5.

[11]Y. Wu, K. Chen, T. Zhang, *et al.*, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP*, 2023, pp. 1–5.

[12]K. Chen, X. Du, B. Zhu, *et al.*, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *ICASSP*, 2022, pp. 646–650.

[13]Y. Liu, M. Ott, N. Goyal, *et al.*, "RoBERTa: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.