3-Q-23

環境音に対する印象説明文データセットの構築と分析

◎ 岡本 悠希（東大院・情報理工），永瀬 亮太郎，岡本 南美（立命館大・情報理工），齋藤 佑樹（東大院・情報理工），福森 隆寛，山下 洋一（立命館大・情報理工）

① 研究背景と目的

環境音 - テキストの相互変換タスクへの注目

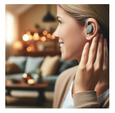
■ Audio captioning: 環境音の内容を自然言語で記述する技術
- 記述例：「猫の鳴き声の後ろで人の笑い声が聞こえる」

■ Text-to-audio: 自然言語から人工的に環境音を合成する技術

環境音 - テキスト相互変換タスクの応用例

■ 映画やゲームなどメディアコンテンツの制作補助，自動生成
- コンテンツに適した環境音を自動生成

■ 聴覚補助
- 補聴器，動画コンテンツへの自動キャプショニング



従来の環境音 - テキスト相互変換タスク

■ 環境音に含まれる内容物（音響イベントの種類など）に焦点

環境音を聴いた際に人間が感じる印象を含むデータが極めて少ない

本研究の目的

環境音の印象情報を記述した説明文データセットの構築
- 印象情報の活用による表現力の高い環境音 - テキスト相互変換を期待

② 従来の環境音 - テキストデータセット

音の内容物・発生順序を記述したデータセット

■ AudioCaps [1]
- 約 5 万個の環境音 - 英語テキストのペアデータが含まれる
- Audio captioning のために設計されたデータセット

■ WavCaps [2]
- 約 40 万個の環境音 - 英語テキストのペアデータが含まれる
- 大規模言語モデルを活用してテキストを作成

課題

人間が環境音を聞いた際に感じる印象情報が含まれていない

本研究

「鋭い」「切ない」のような印象情報のみに着目して
環境音 - テキストのペアデータセットを構築

参考文献

[1] C. D. Kim, et al., “AudioCaps: Generating Captions for Audios in the wild,” Proc. the 2019 conference of the North American Chapter of the Association for Computational Linguistics, pp. 119-132, 2019.
[2] X. Mei, et al., “WavCaps: A ChatGPT-Assited weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 1-15, 2024.
[3] K. J. Piczak, et al., “ESC: Dataset for Environmental Sound Classification,” Proc. ACM Multimedia, pp. 1015-1018, 2015.
[4] Y. Wu, et al., “Large-scale Contrastive Language-Audio Pretraining with Feature and Keyword-to-Caption augmentation,” Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5, 2023.

謝辞

本研究の一部は，JST ムーンショット型研究開発事業 JPMJMS2011，2024 年度国立情報学研究所公募型共同研究 (24S0504)，JSPS 科研費 22KJ3027 及び JST 次世代研究者挑戦的研究プログラム JPMJSP2101 の助成を受けたものです。

③ 印象説明文データセットの構築

構築したデータセットの概要

■ ESC-50 [3] の環境音に対する印象語
- 3,600 個の印象語（30 音響イベント × 40 音 × 3 名）

■ 印象語に対する自信度
- 印象語を付与した本人による 5 段階スコア

■ ESC-50 の環境音に対する印象説明文
- 3,600 個の印象説明文（30 音響イベント × 40 音 × 3 文）
- ChatGPT で生成した後に人手で選別

■ 印象説明文に対する妥当度
- 環境音に対する印象説明文の妥当性を評価した 5 段階スコア

③-A 環境音に対する印象語の収集

クラウドソーシングサービスにより印象語を収集

■ 収集手順
- 被験者に環境音のみを提示
- 被験者は環境音に対する印象を自由記述形式で回答
- 記述した印象語に対する 5 段階の自信度も回答

■ 音に対する印象以外の情報は記述しないように作業者に指示
- e.g.) オノマトペ，音響イベントラベル

MeCab による形態素解析で名詞等を含む回答を除外

■ 形容詞，形容動詞，動詞，形容動詞語幹のみを印象語として扱う
- e.g.) 鋭い，怖い，華やかな

③-B 環境音に対する印象説明文の生成

ChatGPT によって印象説明文の候補を生成

Y

Given sound event and impression word, please generatethe description of sound impression in Japanese

EXAMPLE 1:
Sound Event: ガラスが割れる音
Impression word: 痛々しい
Activate: 緊張した
Valence: 不快
Prohibited word: ガラスが割れる音、緊張した、不快
Description of sound impression: 痛々しい破壊の瞬間を感じさせ、心に落ち着かない音

EXAMPLE 2:
（中略）

INPUT AND OUTPUT
Sound Event: 小鳥のさえずり
Impression word: 鋭い
Activate: 落ち着いた
Valence: 快
Prohibited word: 小鳥のさえずり、落ち着いた、快
Description of sound impression: [output]

INPUT AND OUTPUT
Sound Event: 小鳥のさえずり
Impression word: 鋭い
Activate: 落ち着いた
Valence: 快
Prohibited word: 小鳥のさえずり、落ち着いた、快
Description of sound impression: 鋭く澄んだ音が耳に届き、心穏やかな安らぎをもたらす印象

ChatGPT による印象説明文候補の生成例

■「快 - 不快」，「落ち着いた - 緊張した」の極性情報も入力
- 同じ印象語に対しても様々な極性が想定されるため

クラウドソーシングによって適切な印象説明文を選択

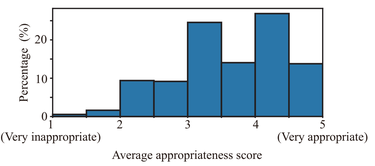
■ 被験者に 4 個の説明文と環境音を提示
- 「快 - 落ち着いた」「快 - 緊張した」「不快 - 落ち着いた」「不快 - 緊張した」の各クラス 100 文生成した後，各クラスごとに 1 文をランダムに選択

■ 被験者は環境音に最も適切な説明文を 1 個選択
- 各音 5 名の被験者によって実施

被験者が最も適切だと回答した印象説明文の例

ChatGPT への入力				出力された説明文
音響イベント	快 or 不快	落ち着いた or 緊張した	印象語	
小鳥のさえずり	快	落ち着いた	鋭い	鋭く澄んだ音が耳に届き、心穏やかな安らぎをもたらす印象
教会の鐘の音	快	緊張した	華やかな	華やかな音色が鳴り響き、新たな始まりを予感させる喜びの音
木のドアを叩く音	不快	落ち着いた	怖い	怖さを感じさせるその重い響きは、心に緊張と不安をもたらす
掃除機の音	不快	緊張した	うるさい	うるさく響き、一瞬で心をざわつかせる雑音

■ 印象説明文の妥当性を評価
- 被験者に印象説明文と環境音を提示
- 被験者は環境音に対する印象説明文の妥当性を 5 段階で評価



印象説明文に対する妥当性スコアの分布

環境音に対して妥当な印象説明文が ChatGPT により生成可能

④ 構築したデータセットの分析

Audio-text retrieval による客観評価

■ 構築したデータセットで CLAP モデルを学習
- CLAP: Contrastive Language-Audio Pretraining [4]
  - 環境音とテキストの埋め込み表現の距離を近づけるようにモデル学習

■ CLAP モデルを用いて audio-text retrieval を実施
- Text-to-audio retrieval (T→A): テキストをクエリとして環境音を検索
- Audio-to-text retrieval (A→T): 環境音をクエリとしてテキストを検索

モデル学習後の評価スコアの向上を確認

■ 環境音と印象説明文のペアはある程度妥当なものであることを確認

Audio-text retrieva の評価結果

Method	A → T				T → A			
	R@1	R@5	R@10	mAP@10	R@1	R@5	R@10	mAP@10
Random[1]	0.003	0.014	0.023	0.007	0.006	0.011	0.034	0.010
Ours[2]	0.034	0.131	0.202	0.077	0.031	0.099	0.168	0.062

[1] Random: モデル学習前のテストデータに対する評価結果
[2] Ours: 構築したデータセットでモデル学習した後のテストデータに対する評価結果

今後の展望：構築したデータセットを用いた環境音 - テキスト変換の実施