S23

# 環境音に対する印象説明文データセットの構築と説明文生成の初期検討

○ 岡本 悠希 [1]，永瀬 亮太郎 [2]，井本 桂右 [3]，山岸 順一 [4]，齋藤 佑樹 [1]，福森 隆寛 [5]，山下 洋一 [2]

[1] 東京大学，[2] 立命館大学，[3] 京都大学，[4] 国立情報学研究所，[5] 滋賀医科大学

## ① 研究概要

### 環境音に対する印象を記述したデータセットを構築

■ 従来のデータセット：音の内容物・発生順序を記述したデータが主

・従来の説明文：「複数の小鳥たちが鳴いている」

・**本研究の説明文：「鋭く澄んだ音が耳に届き、心穏やかな安らぎをもたらす」**

**従来研究の課題**

**人間が音を聞いた際に感じる印象を表す表現が含まれていない**

→ 人間の感性を無視した環境音認識・合成モデルとなる

**本研究**

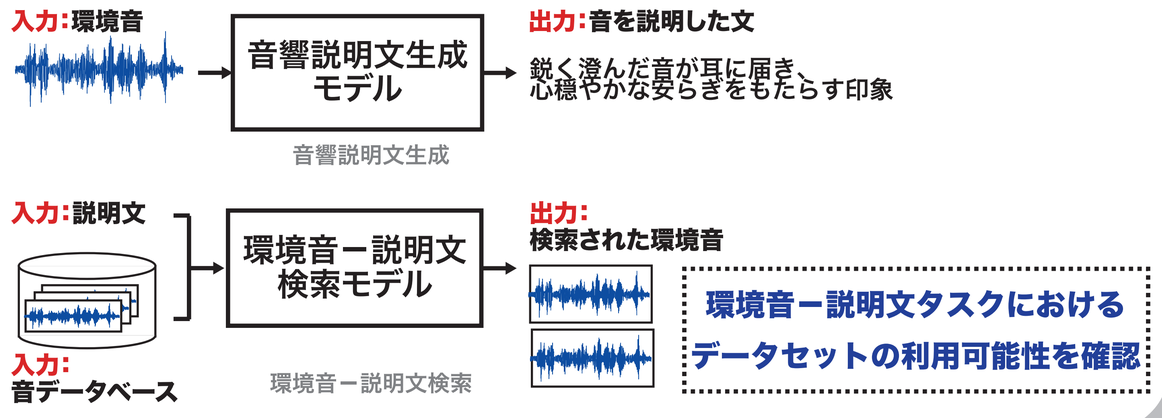
**「鋭い」「切ない」のような印象情報のみに着目して**
**環境音 - テキストのペアデータセットを構築**

■ 応用先：印象情報を活用した表現力の高い環境音認識・合成

### ChatGPT を活用して環境音に対する印象説明文の候補を生成

■ 例：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

### 環境音ー説明文を扱うタスクでデータセットを評価




**環境音ー説明文タスクにおける**
**データセットの利用可能性を確認**

## ② 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 [1] 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

・1（非常に自信がない）～ 5（非常に自信がない）

■ 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）

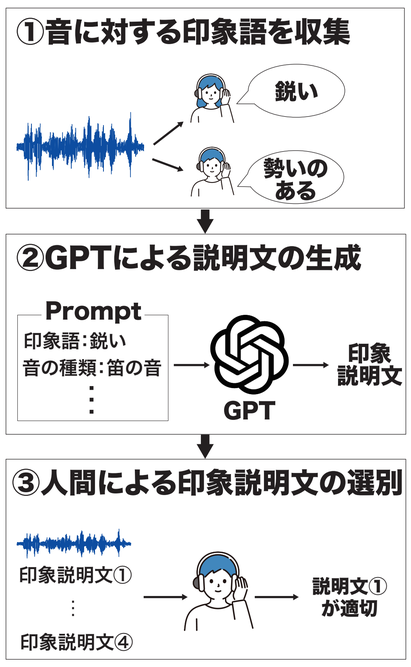
・ChatGPT で生成した後に人手で選別

■ 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

・1（非常に妥当でない）～ 5（非常に妥当である）

### データセット構築の流れ




**① 被験者に音を提示して印象語を収集**

・自由記述形式で収集

・MeCab による形態素解析を実施

・形容詞，形容動詞，動詞，形容動詞語幹のみを印象語とする

**② ChatGPT で印象説明文候補を生成**

・印象語，音の種類，極性情報を入力

・「快 - 不快」「落ち着いた - 緊張した」の極性情報も使用

→ 同じ印象語に対しても様々な極性が想定されるため

**③ 生成した印象説明文候補から人間が選別**

・4 個の説明文と環境音を被験者に提示

・「快 - 落ち着いた」「快 - 緊張した」「不快 - 落ち着いた」「不快 - 緊張した」の各クラスごとに 1 文ずつ選択

・被験者は最も適切だと思う説明文を 1 文選択

### 人間が最も適切だと判断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

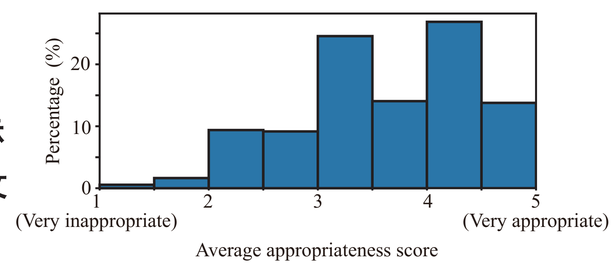
■ 主観評価を実施

・評価手順

・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

・評価者数：各音ー説明文につき 3 名




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**

## ③ 構築したデータセットを用いた環境音ー説明文を扱うタスクでの評価

### 環境音ー説明文検索

■ 構築したデータセットで CLAP モデルを学習

・CLAP: Contrastive Language-Audio Pretraining [2]

- 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 モデルを用いて環境音ー説明文検索を実施

・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であることを確認

環境音ー説明文検索の評価結果

| 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

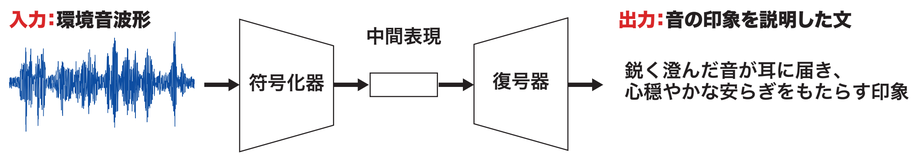
[2] Ours: 構築したデータセットでモデル学習した後のテストデータに対する評価結果

### 音響説明文生成

■ 構築したデータセットで系列変換モデル (seq2seq) を学習

・符号化器：3 層の双方向 LSTM (long short term memory) + 注意機構

・復号器：1 層の単方向 LSTM




### ランダムな生成と比較して評価スコアの向上を確認

■ 意味のある印象説明文生成が可能であることを確認

音響説明文生成の評価結果

| Method | BLUE | | | | METEOR | ROUGE_L | CIDEr | JaSPAICE | BERTScore | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | | | | | Precision | Recall | F1 |
| Random[3] | 0.003 | 0.014 | 0.023 | 0.007 | 0.006 | 0.011 | 0.034 | 0.010 | 62.4 | 59.2 | 60.7 |
| Ours | 0.034 | 0.131 | 0.202 | 0.077 | 0.031 | 0.099 | 0.168 | 0.062 | 74.9 | 73.5 | 74.2 |

[3] Random: データセットに含まれる単語をランダムに並べた文

**今後の展望：構築したデータセットの大規模化 + 環境音生成タスクなどでの評価**

参考文献 [1] K. J. Piczak, et al., "ESC: Dataset for Environmental Sound Classification," Proc. ACM Multimedia, pp. 1015-1018, 2015. [2] Y. Wu, et al., "Large-scale Contrastive Language-Audio Pretraining with Feature and Keyword-to-Caption augmentation," Proc. 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。