音源検索のためのオノマトペを使用した音素材画像化システム

2431145　松浦怜央　　成見研究室

1　はじめに

1.1　研究背景

1.1.1　オノマトペ

オノマトペとは，環境の音や動物の鳴き声・人の叫び声などを模した語である擬音語と，事物の状態を言葉によってそれらしく表す語である擬態語を総称したものである．オノマトペは会話などで頻繁に用いられ，話者が聞き手にその物事の状況をより的確に伝えるために用いられる．

1.1.2　課題

近年，オンライン上には多数のフリー音素材が公開されており，動画制作やアプリ開発などで広く利用されている．これらの音素材はカテゴリ分けやキャプション付与によって整理されているものの，キャプションや説明文から想起される音のイメージと，実際に試聴した音との間に乖離が生じることが多い．その結果，使用者は目的に合致する音素材を見つけるために，多数の候補を一つずつ再生して確認する必要があり，検索に多大な時間的・精神的コストを要している．

1.2　研究目的

本研究では，音素材を入力として与えることでその音素材を表現する画像を作成するシステムを開発する．各音素材ごとにキャプションに加えてその音素材を表す画像（開発したシステムによる出力画像）も表示することによって，使用者の音素材検索に要する時間的・精神的コストが低減されると期待される．システム開発と併せて，本仮説の検証についても実施する．

2　既存研究

2.1　音響信号からの擬音語生成

S. Ikawa らは，Encoder-Decoder モデルを用いて擬音語を自動で生成する手法を提案した [1]．従来手法の課題に対応しつつ，従来手法より低い誤り率を実現できた．本研究では，Ikawa らのアプローチを踏襲し Encoder-Decoder モデルを用いたオノマトペ生成を行う．一方で使用するデータセットは異なる．

2.2　音響信号からのシーン画像生成

Sung et al. は音響信号から適切な環境シーンの画像を合成するモデル「Sound2Scene」を提案した [2]．入力された音声を視覚的特徴に変換し，その後事前学習済みの画像生成器を用いて画像を生成する．本研究とは音そのものを表す画像を生成する点で異なる．

3　開発したシステム

3.1　システム概要

開発したシステムの全体像を図 1 に，音素材の音響的特徴と画像の構成要素の対応関係を表 1 に示す．画像として表現するにあたっては，音素材に含まれる音響的特徴を複数の要素に分解し，それぞれを画像内の構成要素へ対応付けることで視覚化を行う．具体的には，音素材を入力として，音の様相を表すオノマトペおよび音の系統を，学習済みモデルにより推定する．一方で，音の大きさや周波数分布といった音響的特徴は信号処理により解析し，これらの結果を画像内の文字の大きさやフォント，色，背景，および時間平均メル周波数プロファイルとして表現する．

3.2　データセット

国立情報学研究所 音声資源コンソーシアム (NII-SRC) は，特定の環境下で収集した非音声音ドライソース RWCP-SSD を提供している [3]．本研究の学習でこのデータベースの一部（Vol.1）を使用する．

Y. Okamoto らはオノマトペを用いた環境音の合成・変換に関する研究のために，RWCP-SSD Vol.1 に含まれる 105 種類の環境音に対して，計 155,568 個のオノマトペを付与した [4]．GitHub 上で公開されており（RWCP-SSD-Onomatopoeia），利用規約の範囲内に限り無償で利用可能である．

3.3　作成された画像

開発システムによって作成された画像を図 2 に示す．

4　評価実験

開発したシステムの有効性を検証するため，以下の評価実験を行った．

1. 音素材のキャプション評価アンケート
2. Web テスト
3. 事後アンケート
4. 内省報告（2, 3 を早く終えた人のみ）

Web テストでは，「開発システム画像 + キャプション」「画像なし（キャプションのみ）」「ChatGPT 生成画像 + キャプション」の 3 条件（それぞれ X1, X2, X3）に対し，同一音素材選択問題（問題形式 B）およびシーン適合音素材選択問題[*1]（問題形式 C）の回答および回答時間の収集を行った．

[*1] 回答者に音素材の利用シーンを文字で提示し，そのシーンにあった音素材を選択肢から選んでもらう問題

5 結果

本研究におけるデータは，倫理審査を行った上で，電気通信大学の学部生および大学院生の男女計 33 名から収集した．

5.1 開発したシステムでの回答時間と完遂率

B 問題の回答時間に対し，ペナルティを課した「タスク完了時間」の分析を行った結果を表 2 に示す．選択肢数が 40 個の場合（Y2）において開発システム条件（X1）が特に優位であり，選択肢数が 20 個の場合（Y1）と同等の回答時間でタスクの完了が確認でき，なおかつ画像なし条件（X2）と比較し回答時間が有意に短かった（$p < .05$）．

また完遂率については，Y1, Y2, Y3 すべての場合において開発システム条件（X1）と画像なし条件（X2）との間に有意な差は認められなかったが，選択肢数が 40 個の場合（Y2）では画像なし条件（X2）と比較して 15.1 ポイント高い 90.9% と，比較的高い水準であった．

問題形式 C においても同様の分析を行った．その結果，開発システム条件（X1）と画像なし条件（X2）との間で，回答時間に有意な差は認められなかった．完遂率についても，Y1, Y2, Y3 すべての場合において開発システム条件（X1）と画像なし条件（X2）との間に有意な差は認められなかったが，選択肢数が 40 個の場合（Y2）では画像なし条件（X2）と比較して 15.1 ポイント高い 72.7% と，比較的高い水準であった．

5.2 開発したシステムでの検索の容易さとストレス/疲労

事後アンケートの結果に対し，提示方式を要因とした参加者内一元配置分散分析を行った結果を表 3 に示す．検索の容易さおよびストレス/疲労に関して，開発システム条件（X1）が他の 2 条件より有意に低かった（$p < .05$）．つまり，開発したシステムによる画像をキャプションと併せて表示することで，検索が容易となり，検索によるストレス/疲労を感じにくくなる事実が明らかとなった．

6 まとめと今後の課題

6.1 まとめ

音素材ごとに，キャプションとその音素材を適切に表す画像を表示することで，

- **検索に要する時間的コストは低減する場合がある**（選択肢数，タスクにより異なる）
- **検索に要する精神的コストは低減する**

6.2 今後の課題

今後の課題として，未知の音への対応，オノマトペの整合性向上，時間平均メル周波数プロファイルおよび音高の代表値の有効性検証などが挙げられる．

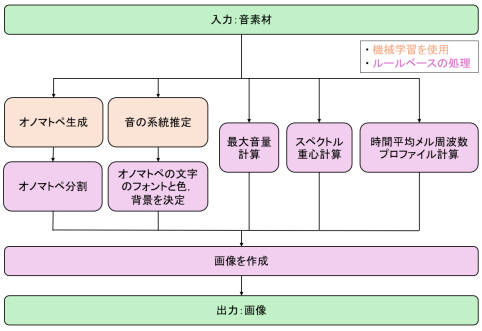

図 1 システムの全体像

表 1 音素材の音響的特徴と画像の構成要素との対応関係

音素材の音響的特徴	画像の構成要素
音の様相	オノマトペ
音の大きさの最大値	オノマトペの文字の大きさ
音の系統	オノマトペの文字のフォントおよび色，背景
音の高さの分布	ヒートマップ「時間平均メル周波数プロファイル」
音の高さの代表値	ヒートマップへ向けた「▼」
音の長さ	オノマトペの長さ

図 2 開発したシステムで作成された画像（4 つを抜粋）

表 2 問題形式 B における提示方式と選択肢数による回答時間の比較（$N = 33$）

提示方式	20 個 (Y1)	40 個 (Y2)	100 個 (Y3)
開発システム (X1)	32.83 ± 22.91	32.69 ± 18.49	78.85 ± 39.97
画像なし (X2)	20.43 ± 12.67	49.82 ± 25.47	72.35 ± 41.48
ChatGPT (X3)	47.63 ± 25.13	60.13 ± 27.15	111.05 ± 21.27

数値は平均値 ± 標準偏差（秒）．タイムアウトおよび不正解は 120 秒として処理

表 3 提示方式によるユーザー評価の比較（$N = 30$）

評価項目	画像なし	開発システム	ChatGPT
検索の容易さ（高が良い）	2.77 ± 1.17	4.07 ± 0.68	2.27 ± 0.89
ストレス・疲労（低が良い）	3.17 ± 1.16	2.33 ± 0.98	4.07 ± 0.93
回答への自信（高が良い）	3.63 ± 1.02	4.03 ± 0.71	3.03 ± 1.11

数値は平均値 ± 標準偏差

参考文献

[1] S. Ikawa and K. Kashino. Generating sound words from audio signals of acoustic events with sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 346–350, Calgary, AB, Canada, 2018.

[2] Sung-Bin Kim, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. Sound to visual scene generation by audio-to-visual latent alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6430–6440, 2023.

[3] S. Nakamura, K. Hiyane, F. Asano, and T. Endo. Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pp. 965–968, 2000.

[4] Yuki Okamoto, Keisuke Imoto, Shinnosuke Takamichi, Ryosuke Yamanishi, Takahiro Fukumori, and Yoichi Yamashita. Rwcp-ssd-onomatopoeia: Onomatopoeic word dataset for environmental sound synthesis. *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 125–129, 2020.