大規模言語モデルを用いたオノマトペ付与による日本語音声データセットの拡張 [S1-P21]

小川 剛毅[1], 根本 颯汰[2], 北田 俊輔[2], 彌冨 仁[1,2] [1]法政大学 理工学部, [2]法政大学大学院 理工学研究科

Summary

ChatGPTを用いたオノマトペ付与による日本語音声データセットの拡張手法の提案

- 大規模言語モデル (LLM) を用いて既存データセットにオノマトペを付与するデータ拡張方法の提案
- 提案手法により拡張した Extended-Audiocaps データにより text-to-audio タスクで性能向上を確認

Background

- 日本語における LLM を利用した音声認識の問題点
  - 😭 データ不足・英語の直訳テキストの使用
- 特に英語データに対する機械翻訳を用いる弊害
- 昨今、感性を表現するオノマトペは言語学、音声学どちらにおいても注目を集めている
  - 😭 英語から直訳するとオノマトペが不自然のまま
  - 😭 オノマトペの有無が検索精度に依存

➔ **オノマトペの改善や付与を行い検索精度向上に向けた初期検討を行う**

Proposed Method

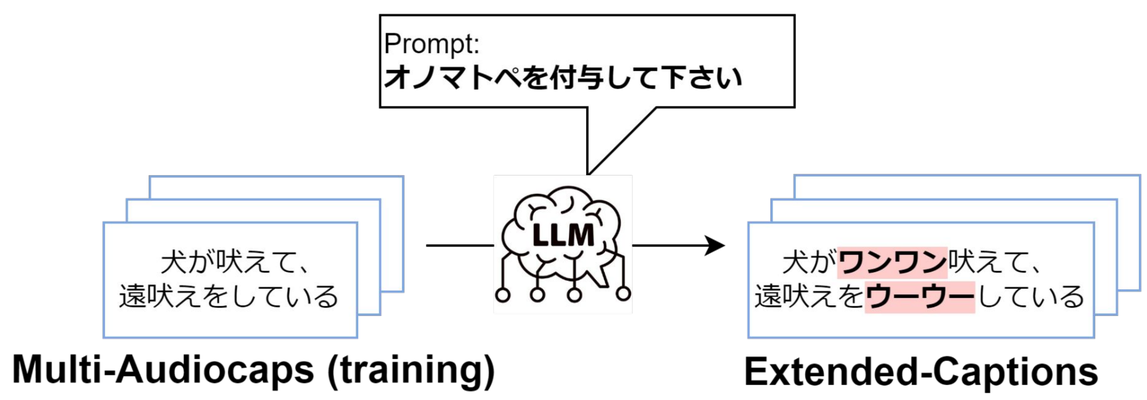

Fig.1 LLMオノマトペデータ生成の概要図

LLM によるオノマトペデータの取得方法

- LLM に音声データに含まれるキャプションデータを指示テキストと共にプロンプトとして入力
- 得られた結果を **Extended-Audiocaps** として使用

Multi-Audiocaps に対する適用

- テキストと対応する音声が含まれる Audiocaps [Kim+,NAACL] を機械翻訳した Multi-Audiocaps [岡本+] を拡張
- 拡張時に LLM として gpt-4o-mini を使用

Table.1 ChatGPT を用いたオノマトペ付与結果

Dataset	Example
🇺🇸 Audiocaps [Kim+,NAACL]	A dog is howling and barking.
🇯🇵 Multi-Audiocaps [岡本+]	犬が吠えて、遠吠えをしています。
🇯🇵 **Extended-Audiocaps (ours)**	犬が**ワンワン**吠えて、遠吠えを**ウーウー**しています。

Experiments & Results

実験設定

- CLAP (Contrastive Language-Audio Pretraining) [Elizadle+ ICASSP'23] を Extended-Audiocaps で fine-tuning

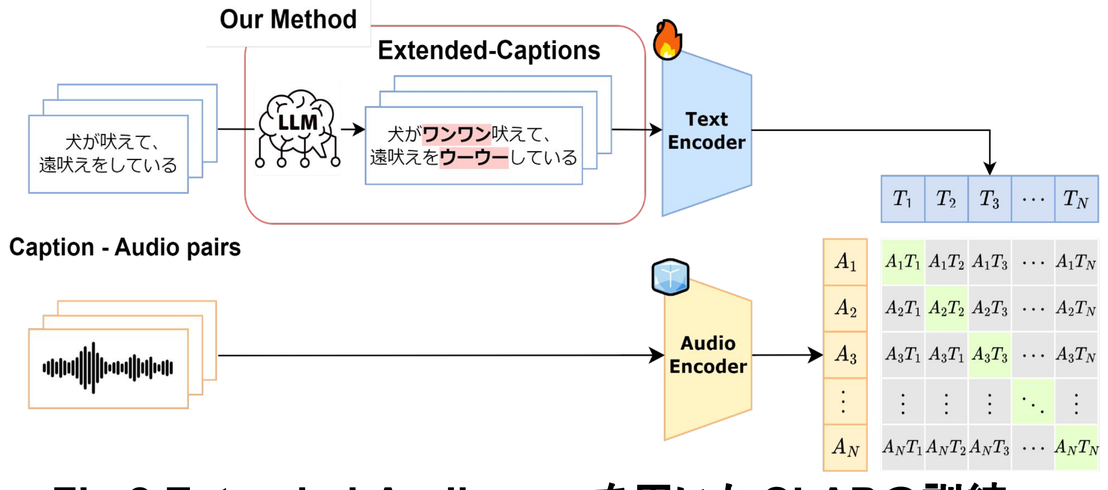

Fig.2 Extended-Audiocapsを用いたCLAPの訓練

- Text-to-Audio 検索タスクを用いて評価を実施
  - 入力として与えるテキストの埋め込みを用いて類似する音声データを検索する性能で評価
    - Recall@k: cosine類似度Top kの再現性
    - mAP@k: cosine類似度Top kまでの適合率の平均

実験結果

Table.2 Multi-AudiocapsとExtended-Audiocapsの比較

Dataset	Recall@1	Recall@5	Recall@10	mAP@10
Multi-Audiocaps [岡本+]	0.249	0.577	0.713	0.383
Extended-Audiocaps	**0.254**	**0.583**	**0.720**	**0.388**

😊 **提案手法により両方の評価指標でスコア向上**

Discussion & Future Work

Discussion

- 🤔 全データが精度向上に影響しているか不明
  - ➔ 有効なデータの特定と拡張
- 😭 実際のオノマトペと乖離しているデータが存在
  - ex. ***“奥で微かに「こもれび」と聞こえる流れる「さらさら」とした水”***
  - ➔ プロンプト改善により不自然なデータ生成を防止

Future Work

- Data augmentation 手法としての適用
- オノマトペ付与のためのプロンプトおよび精度の調査
- 日本語以外の言語に対する拡張と有効性の検証

Acknowledgement

東京大 岡本悠希 先生並びに慶応大 高道慎之介 先生には本研究の遂行にあたり多大なご協力頂きました。
この場をお借りして深く感謝申し上げます。