S3-P12

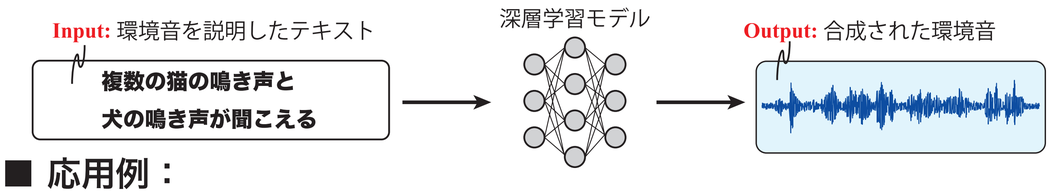
Text-to-audio における評価指標 CLAP-Score の性能分析

高野 大成，岡本 悠希，齋藤 佑樹（東京大学）

① 研究背景・目的

Text-to-audio とは

■ テキストを入力として環境音を合成する技術



■ 応用例：

・映画などのメディアコンテンツ作品における効果音・背景音の自動生成

Text-to-audio の評価

■ CLAP-Score によりテキストと合成音の対応関係を評価

・入力テキスト内の情報が合成音に反映できているか評価

■ 課題：

・**入力テキストの時系列関係が適切に評価できているか不明**

e.g.)「猫の鳴き声の後に車のクラクションが聞こえる」

・**人間による評価との相関関係が明らかでない**

本研究の目的

CLAP-Score の評価性能の分析

・テキスト内の時系列関係と合成音との対応は適切に評価できているか

・人間による主観評価との相関関係はあるか

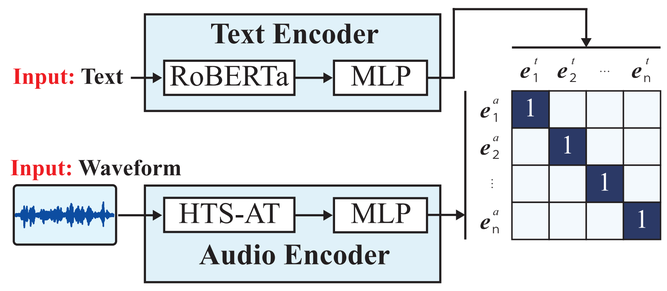
② CLAP-Score とは

CLAP を用いてテキストと環境音の対応関係を評価

■ CLAP: Contrastive Language-Audio Pretraining [1]

・**ペアとなる音とテキストの埋め込みベクトルを近づけるようにモデル学習**

・画像分野で使用される CLIP の音バージョン



Text encoder

・テキストから埋め込みベクトルを獲得

・学習済みの RoBERTa [2] などのモデルを使用

Audio encoder

・音波形から埋め込みベクトルを獲得

・学習済みの HTS-AT [3] や PANNs [4] を使用

■ CLAP-Score の定式化

・**CLAP から得たテキストと音の埋め込みベクトル間のコサイン類似度を計算**

・スコアが大きいほどテキストと音の対応が取れていると評価

$$\text{CLAP-Score}(e^a, e^t) = \frac{e^a \cdot e^t}{\|e^a\|\|e^t\|}$$

e^t：テキストの埋め込みベクトル

e^a：音の埋め込みベクトル

③ CLAP-Score の時系列性能の評価

作成した擬似データに対して CLAP-Score を計算

1. テキストと音データの CLAP-Score を**そのまま**計算
2. 擬似データの**時系列を入れ替えて** CLAP-Score を再計算

・テキストのみ入れ替え / 音データのみ入れ替えの両パターン計算

・**CLAP-Score が時系列を評価できていればスコアは減少するはず**

時系列を含むテキスト - 環境音の擬似データ作成

■ 使用したデータセット：ESC-50 [5]

・様々な種類の環境音 (e.g. 犬，水滴）を集めたデータセット

■ 二つの環境音が連続して発生するように音データを結合

■ 結合した環境音に対するテキストを作成

e.g.) “Sound of dog barking followed by sound of engine running”

CLAP-Score と主観評価の相関を分析

■ 入れ替え前後のスコア変化率を計算

$$\text{スコア変化率} = \frac{\text{CLAP-Score(時系列入れ替え)}}{\text{CLAP-Score(時系列そのまま)}}$$

■ 時系列を入れ替えても CLAP-Score に変化なし

→ **CLAP-Score は時系列を評価できる客観的指標ではない**

擬似データに対する CLAP-Score とスコア変化率

環境音の種類	CLAP-Score（時系列そのまま）↑	CLAP-Score（時系列入れ替え）↑	スコア変化率
犬 - 猫	0.2332	0.2340	1.00
犬 - エンジン	0.1790	0.1790	1.00
犬 - 水滴	0.4042	0.4089	1.01

④ CLAP-Score と主観評価の比較分析

合成音に対して CLAP-Score と主観評価結果を比較

■ CLAP-Score が**人間のどの主観評価と相関があるかを分析**

■ 2 種類の主観評価（5 段階評価）

・テキストに記述された音が合成音にどの程度含まれるか (Inclusion Score)

→ 1 (Clearly not Included) ～ 5 (Clearly Included)

・テキストの時系列と合成音がどれだけ一致しているか（Order Score)

→ 1 (Very Poor) ～ 5 (Very Good)

Text-to-audio を用いて時系列を含む環境音を合成

■ 使用したデータセット：AudioCaps [6]

・テキスト - 環境音のペアから構成されるデータセット

■ 時系列を含むテキストを用いて環境音を合成

・テキストは AudioCaps から抽出（100 サンプル）

・使用モデル：AudioLDM [7], AudioLDM2 [8], Tango [9], Tango2 [10]

→ **合成音 400 音（各手法につき 100 音）+ AudioCaps の自然音 100 音を評価**

クラウドソーシングにて主観評価実験を実施

■プラットフォーム：Prolific

■ 被験者：英語話者 74 名

CLAP-Score と主観評価の相関を分析

■ CLAP-Score と主観評価結果の相関係数

・Order Score: 0.512

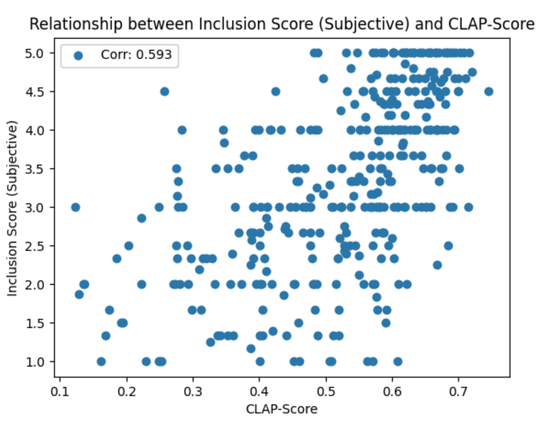
・Inclusion Score: 0.593

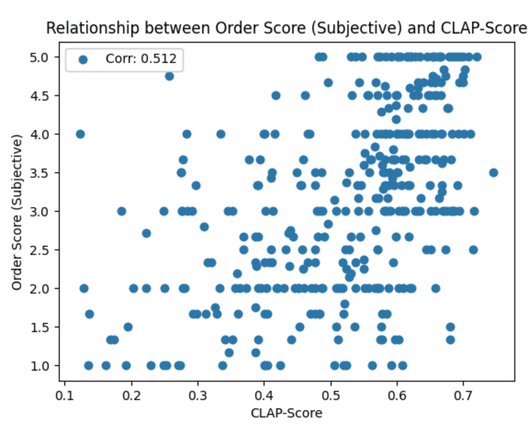
どちらも相関が認められるが強いものではない

CLAP-Score と主観評価の関係

■ 主観評価と CLAP-Score の相関が低いデータを個別に分析

→ 時系列よりもテキストの複雑さや音の欠落等が CLAP-Score に影響





CLAP-Score は主観評価を代用できる客観的指標としては不十分

参考文献

[1] Y. Wu, et al., Proc. ICASSP, 2023.
[2] Y. Wu , et al., arXiv preprint, 2019.
[3] K. Chen et al., Proc. ICASSP, 2022.
[4] Q. Kong, et al., IEEE/ACM TASLP, 2022.
[5] K. J. Piczak, et al., Proc. ACM Multimedia, 2015.
[6] C. D. Kim, et al., Proc. NAACL, 2019.
[7] H. Liu, et al., Proc. ICML, 2023.
[8] H. Liu, et al., IEEE/ACM TASLP, 2024.
[9] D. Ghosal et al., arXiv preprint, 2023.
[10] N. Majumder, et al., Proc. ACM Multimedia, 2024.

今後の展望

■ CLAP-Score と人間の主観との関係についてのさらなる調査

■ 時系列を適切に評価できる新たな客観的指標の策定

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 及び，2024 年度国立情報学研究所公募型共同研究 (24S0504) の助成を受けたものです