TSUBAME 共同利用　令和7年度　学術利用　成果報告書

利用課題名　生成型音声・画像 AI のための学習アルゴリズムの開拓と自動機械学習の研究
英文：Investigation of Training Algorithms and AutoML for Generative AI

利用課題責任者: 坂東 宜昭
First name Surname: Yoshiaki Bando

所属: 国立研究開発法人　産業技術総合研究所
Affiliation: National Institute of Advanced Industrial Science and Technology (AIST), Japan
URL: https://www.airc.aist.go.jp/cosine

邦文抄録（300 字程度）

本利用課題では，実世界で人間と協調する次世代人工知能（AI）技術を確立するため，その効率的な学習方法の開拓とともに，実応用で課題となる自動機械学習（AutoML/AutoPEFT）に取り組んだ．特に，音響イベント検出（SED）において，テキストプロンプトに基づき任意の音響イベントの発生時刻を推定する手法の開発に取り組んだ．具体的には，テキスト埋め込みモデルを導入し，環境音の埋め込みとイベント名の埋め込み表現との関係を大規模に学習することで，環境音埋め込みからテキストプロンプトに対応する音響イベントをゼロショットで検出する枠組みを構築した．さらに，このような研究の基盤となる，高性能計算（HPC）と親和性の高いマルチモーダル環境音分析のためのソフトウェア基盤を整備した．

英文抄録（100 words 程度）

We addressed efficient training methods to establish artificial intelligence (AI) technologies that can collaborate with humans in real-world. Specifically, we focused on developing sound event detection (SED) that estimates the temporal activations of arbitrary sound events based on text prompts. We built a multi-modal framework that performs zero-shot detection of sound events directly from audio embeddings. We introduced a text embedding model to learn the relationship between audio embeddings and label-name embeddings at scale. Furthermore, we developed a software infrastructure for multi-modal acoustic scene analysis that is highly compatible with high-performance computing (HPC). We believe that our contribution serves as the foundation for this line of research.

Keywords: Multi-modal training, generative AI, contrastive learning, audio language model, text-prompting,

背景と目的

大規模にスケールする言語モデルや自己教師あり学習の台頭により，多くの人工知能（AI）タスクが実用に足る性能を達成している．一方，我々の日常生活で人々とコミュニケーションをとりながら適切に仕事をこなす技術は未だ課題が残る．本研究課題の目的は，このような人間と協働する AI 技術を確立することである．特に，ユーザの利用環境に応じて適応的に高い性能を出すための枠組み，ユーザの要望に応じて適宜自身を変化させることができるメタ学習技術の確立を目指した．

概要

本利用課題では，実世界で人間と協調する AI 技術を確立するため，その効率的な学習方法の開拓とともに，実応用で課題となる自動機械学習に取り組んだ．具体的には，テキストプロンプトに基づき任意の音響イベントの発生時刻を推定する音響イベント検出（SED）の開拓や，マルチモーダル環境音分析のためのソフトウェア整備など，実世界を理解し人々と協調する AI に必要な要素技術の開拓を進めた．

結果および考察

従来の SED では，事前に定義した所望の検出対象の音響イベントのみを検出する手法が一般的であり，定義されていないイベントの検出には再学習が必要であった．しかし，実用上はあらゆる環境音を目的に合わせて検出するために，任意の種類の音響イベントを指定して検出できる枠組みが望ましい．そこで我々は，テキスト埋め込みモデルを導入し，環境音の埋め込みとイベント名の埋め込み表現との関

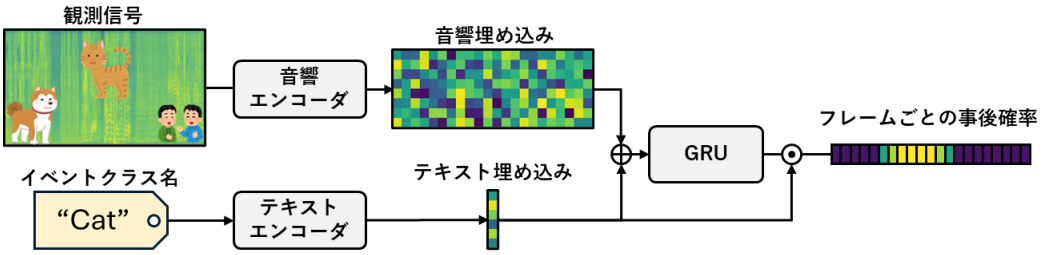

図 1 構築したプロンプト可能な SED システムの概要図．音響エンコーダには BEATs を用いた．

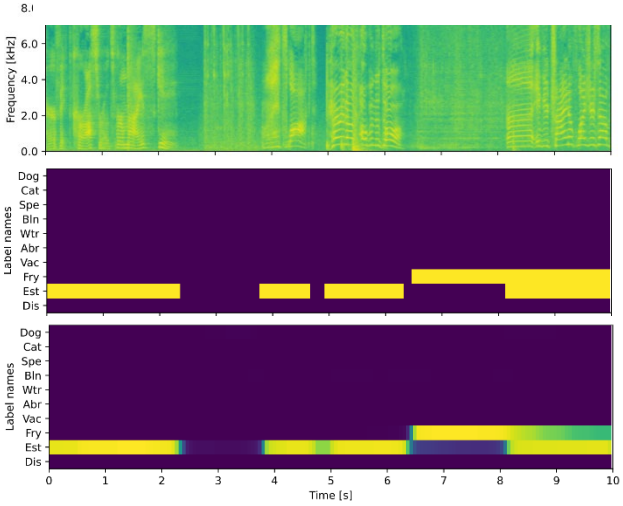

図 2 構築した SED システムの推定結果（下段）．上段と中断は入力信号と正解ラベルを示す．

係を大規模に学習することで，環境音埋め込みからテキストプロンプトに対応する音響イベントをゼロショットで検出する枠組みを構築した．

構築したモデルでは，テキスト埋め込みモデルに RoBERTa を使用し，音響埋め込みモデルには BEATs を用いた（図 1）．本モデルは AudioSet の強アノテーションデータを用いて事前学習したのち，屋内環境音のベンチマークである DESED にて有効性を確認した（図 2）． 特に，図 3 に示す通り，ラベル名を事前に正則化しておくこと，クリップごとに発生していないクラスラベルに重みづけすること（負例重み）が性能改善に重要であることを確認した [1]．

さらに， HPC 環境向けの基盤ライブラリである aiaccel（昨年度採択課題の成果）を拡張するとともに，マルチモーダル環境音分析ソフトウェア Multi-Modal Machine Listening (M3L) Toolkit の開発を進めた．本ソフトウェアは，膨大な学習データを HPC 上で容易にハンドルしながら環境音分析モデルを学習・推論するためのツールキットである．例えば，AudioSet は 6 百万サンプルの音響クリップからなり，非圧縮で約 1TB の容量となる．さらに，これらのデータはウェブ上の動

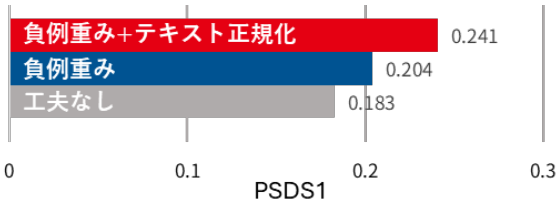

図 3 提案法の音響イベント検出性能（PSDS1）．

画から構成されるため，学習には `ffmpeg` や `sox` 等を用いた前処理が不可欠である．本研究課題では，これらのデータを効率的に前処理するため，ジョブスケジューラを抽象化した `aiaccel.job` を実装した．

`aiaccel.job` では， Slurm や PBS に加えて，TSUBAME 4.0 を想定した SGE 用の実装を追加した．これにより，計算機環境を殆ど意識せず， 単一のコードを用いて複数環境で画一的に動作するソフトウェアの基盤を整備した．同様のソフトウェアに Dask や Parsl があるが，`aiaccel.job` はより軽量な抽象化として実装されている． ESPnet や Kaldi を参考とし，管理サーバを用いず，シェルベースのジョブディスパッチャーとして実装した．これらが Perl で実装されていたのに対し，`aiaccel.job` では，保守性を考慮して YAML 設定ファイルを用いる Python 実装とした．

まとめ、今後の課題

本利用課題では，実世界で人間と協調する次世代人工知能（AI）技術を確立するため，その効率的な学習方法の開拓とともに，実応用で課題となる自動機械学習に取り組んだ．特に，音響イベント検出（SED）において，テキストプロンプトに基づき任意の音響イベントの発生時刻を推定する手法の開発を実施した．

参考文献

[1] 音響・テキストマルチモーダル学習に基づくプロンプト可能な音響イベント検出，神取 雄大，櫻井 舜，坂東 宜昭，井本 桂右，大西 正輝，音学シンポジウム，pp. 69–72, 2025.