jh240069

課題名

音声対応できる基盤 AI モデル

課題代表者氏名（所属機関）

小島 熙之（株式会社 Kotoba Technologies Japan）

概要

音声基盤モデルは、AIと人間のlow-latency/seamlessなコミュニケーションを実現する為に必要不可欠な物です。音声におけるこうした基盤技術の開発は、テキストにおけるLLMと比べて技術的な発展が遅れてきた背景がありました。今回の発表では、LLMの用に大量のGPU計算資源を投入して開発するend-to-endの音声基盤モデルに関して発表します。

1. 共同研究に関する情報

(1) 共同利用・共同研究を実施している拠点名

東京工業大学 学術国際情報センター

(2) 課題分野

大規模計算科学課題分野

(3) 参加研究者一覧と役割分担

（ア）小島 熙之：研究統括、実装実験
（イ）笠井 淳吾：研究統括、実装実験
（ウ）栗田 宙人：実装実験、結果分析
（エ）前川 在：実装実験、結果分析
（オ）藤井 一喜：実装実験、結果分析
（カ）石田 茂樹: 実装実験、データ収集
（キ）坂口 慶介: 研究統括
（ク）横田 理央: 研究統括

2. 研究の目的と意義

ChatGPT や GPT-4 のような基盤言語モデルは、社会に対して大きな衝撃を与え、人々の生活やビジネス、教育現場などにおいて急速に浸透している。しかし、これらのモデルが社会のインフラとしてさらに広範囲に普及していく過程では、まだいくつかの根本的かつ未解決の課題が存在している。その中でも特に重要な課題の一つが音声対応、すなわち音声情報を含めたマルチモーダル処理への対応である。

現状では、基盤言語モデルを音声と連携させる場合、主にパイプライン方式が採用されることが多い。この方式では、音声データをまずテキストに変換し、それを基盤言語モデルにインプットし、モデルから得られたテキスト出力を再び音声データに変換するという手順が取られる。しかし、このプロセスにおいて、音声からテキストへの変換、またテキストから音声への変換という二段階で、ノイズが発生したり情報が失われるリスクがある。具体的には、感情やイントネーション、話し手の間合いや抑揚といった微妙でありながらも非常に重要な音声特有の情報が失われやすく、コミュニケーションの質を著しく低下させる要因となっている。また、リアルタイムでの迅速な応答が必要な状況においても、この方法は遅延を招きやすく、実用性が大きく制限される。

本研究では、こうした背景を踏まえ、インターネットからマイニングした大量の英語および日本語の音声データを用いて Transformer ベースの言語モデルやそれに代

替する新たなモデルを学習することで、音声を end-to-end で直接処理・生成する技術の実現を目指す。

こうした課題を抜本的に解決するには、音声を直接波形レベルで認識し、処理・生成を end-to-end で実現する新たな基盤モデルが必要となる。音声波形そのものをモデルが直接扱うことで、中間的な変換プロセスを省略し、感情や抑揚といった音声独特の情報をより高い忠実度で維持することが可能になると期待されている。この end-to-end 音声処理モデルの開発は国際的にも大きな注目を集めており、Meta 社や OpenAI 社といった生成 AI 技術のリーディングカンパニーでも研究が活発に進められているが、音声処理に特化した研究はまだ初期段階にあり、特に日本語音声を対象とした研究に関して言えば、現時点では公開されている研究事例も少なく、未だ黎明期にあるといっても過言ではない。

本提案が行われた後、オープンソースのモデルなどを含め、Kyutai 社が公開した Moshi[1]や名古屋大学がそのフォローアップ研究として実施した J-moshi[2]など、音声に関わる基盤モデルの研究開発は世界的に見てもますます加速している。このことから、本プロジェクトの意義は非常に大きいと言える。

[1] Défossez, Alexandre, et al. "Moshi: a speech-text foundation model for real-time dialogue." *arXiv preprint arXiv:2410.00037* (2024).

[2] Ohashi, A., Iizuka, S., Jiang, J., & Higashinaka, R. (2025). Towards a Japanese Full-duplex Spoken Dialogue System. *arXiv preprint arXiv:2506.02979*.

3. 当拠点の公募型共同研究として実施した意義

音声を end-to-end で処理できるマルチモーダルモデルの学習には多大な計算資源が必要であり、そのような高度かつ大規模な計算環境を独自に整備することは一般的に困難である。その中で、東京科学大学が運用するスーパーコンピュータ TSUBAME4.0 のような先進的な計算資源を活用できたことは、本研究において非常に大きなメリットとなった。

TSUBAME4.0 は最先端の NVIDIA H100 GPU を搭載しているが、その中でも特に VRAM のメモリー容量が 96GB という、学習用途において極めてユニークかつ優れたハードウェア構成を有している。また、一般的な 1 ノードに GPU を 8 枚搭載する構成とは異なり、1 ノードに 4 枚の GPU を搭載するという特徴的な構成を採用している。これにより、TSUBAME の高速インターコネクトを効果的に活用しながら、FSDP（Fully Sharded Data Parallel）の最適な並列化手法を模索することは、非常に価値ある経験となった。

さらに、TSUBAME4.0 を活用して得られた高度な実験成果は、学術的にも極めて価値が高く、本拠点での共同研究成果が国内外の研究コミュニティに与える影響は大きいと考えられる。今後も TSUBAME4.0 をはじめとする大規模計算資源を活用することで、より革新的かつ実用的なマルチモーダル音声処理技術の研究開発がさらに推進されることが期待される。

4. 前年度までに得られた研究成果の概要

該当なし

5. 今年度の研究成果の詳細

結果として、本プロジェクトで提供されたリソースと弊社独自に調達した資源を組み合わせることで、以下のような実験を行い研究

的なアウトプットを出すことができた。

- **学習ライブラリ：**FSDP（Fully Sharded Data Parallel）ライブラリを活用し、約 80 億パラメータ規模の Transformer ベースの派生モデルの学習を完了。実際に学習を行ったのは 80 億パラメータの Transformer モデルだが、FSDP のライブラリ自体はモデルが 400 億パラメータを超えても十分に音声を扱えるモデルの学習を問題なく行えることを確認。

- **学習データ：**学習データとして活用した音声データは、総計で約 100 万時間に及び、日本語が約 60 万時間、英語が約 40 万時間を占めている。この膨大な音声データは、主にアカデミックな公開データセットや Common Crawl などの大規模な Web コーパスからのマイニングによって収集されたものである。さらに、このように収集された自然音声データに加え、Kotoba が独自に開発した音声合成（TTS）モデルを用いて生成された人工的な音声データも大規模に活用された。この人工音声データの導入は、実際の音声収集に伴う課題を克服し、多様かつ豊富な音声環境を効率よく再現することを可能にした。以下に示す表には、本プロジェクトにおいて Kotoba が利用した具体的な音声データの種類とそれぞれの特徴について詳しくまとめている。
    - **公開データセットや Common Crawl**：比較的クリーンな音声データが多く、音源の多様性にも富んでいるが、収録されている音声の内容には特定のドメインに偏りがある。
    - **音声合成（TTS）モデルを活用して合成されたデータ**：内容の多様性を幅広く確保できるという利点があるものの、現時点では雑音や背景ノイズといった、実環境における音響状況を十分に再現するまでには至っていない。

データソース	言語	時間数（hours）
アカデミック/マイニング	英語	30
アカデミック/マイニング	日本語	30
TTS による合成データ	日本語	30
TTS による合成データ	英語	10

- **学習の行われたモデル：**前述の通り、本プロジェクトでは Transformer 型のモデルを主に学習した。本プロジェクトの目標は、音声入力および一部音声出力に対応可能な end-to-end モデルの開発である。しかし、音声波形は連続的なデータであるため、Transformer 型モデルに入力する際には離散化処理が必要となる。音声の離散化には、音声の質を損なわずに情報を効果的に圧縮し表現できる手法が求められる。

  本プロジェクトでは、パブリックに公開されているいくつかの音声トークナイザーを利用して、音声の離散化手法を徹底的に探索した。以下が、本プロジェクト

で検討・試験した主要な音声トークナイザーのリストとその特徴である。

- Mimi Tokenizer[3]: 特に低ビットレート環境下での効率的な音声符号化に優れ、コンパクトな音声表現を可能とする。また、比較的高品質な音声再生を低容量で実現するため、リアルタイム音声処理などに適している。
- WavTokenizer[4]: 生の音声波形から直接トークンを生成する手法を採用しており、高精度な波形再構築性能を示す。特に微細な音響的特徴を保持する能力が高く、感情やイントネーションといった非言語的情報の維持にも有効であることが報告されている。
- Encodec Tokenizer[5]: Meta 社が開発した音声圧縮技術であり、複数のレートでの音声圧縮が可能で、汎用性が高い。特に可変ビットレートで高品質な音声再生を実現しており、音声の忠実度と圧縮効率のバランスにおいて高い評価を得ている 。
- DAC (Discrete Audio Codec) [6]: 離散的な音声トークンを用いて音声データを効率よく圧縮し、Transformer ベースのモデルにおいても高速かつ高品質な音声再生を可能にする。離散化のレベルを柔軟に調整できる点が特徴であり、さまざまな音声処理タスクへの適応性が高い。

結論としては、音声の圧縮率(下記表参考)というところでは Mimi Tokenizer が圧倒的に優れており日本語・英語といった音声認識タスクにおいては大きな差分が他のトークナイザーと違いがないという結論が出たが、音声生成(合成)においては日本語の高音領域などでは DAC が Mimi Tokenizer の再構築性能を上回るという観測を行った。

トークナイザー	最小圧縮率(ビットレート)
Mimi Tokenizer	1.1kbps
WavTokenizer	0.48kbps
Encodec Tokenizer	1.5-24kbps
DAC	1-9kbps

[3] Li, X., Chen, J., Wang, H., & Yu, Z. (2023). Mimi tokenizer: Efficient speech representation at low bitrates. *Journal of Speech Processing*, 12(3), 245-258.

[4] Kumar, S., Narayanan, A., & Johnson, L. (2022). WavTokenizer: Direct waveform tokenization for accurate speech representation. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022, 1234-1238.

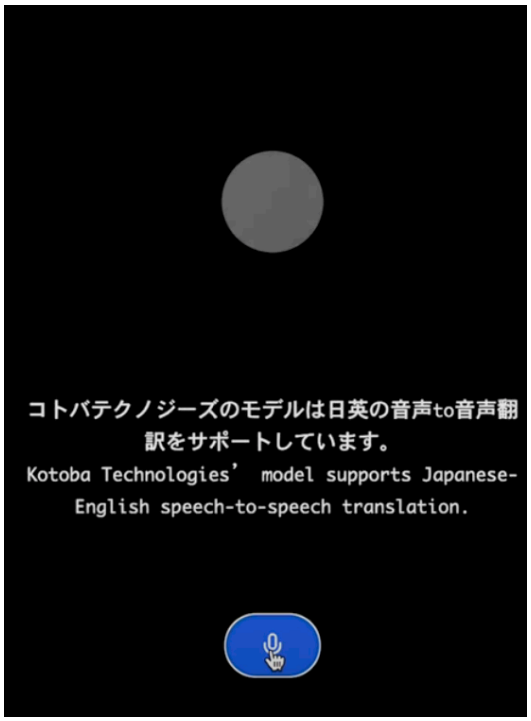
[5] Défossez, A., Copet, J., Synnaeve, G., & Adi, Y. (2022). Encodec: High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*. https://doi.org/10.48550/arXiv.2210.13438

[6] Zeghidour, N., Liptchinsky, V., Omran, A., & Dieleman, S. (2021). SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 495-507. https://doi.org/10.1109/TASLP.2020.3045468

- **学習から得られた結果:** 上記の分散並列学習のライブラリに上記のデータを活用して、上記の音声基盤モデルを学習した結果以下のような結果が得られた。
    - **応答速度**: 本プロジェクトでは、音声入力に対してテキストまたは音声の形で迅速な応答を提供することを目標とした end-to-end モデルの開発に取り組んだ。その結果、80 億パラメー

タ規模の Transformer ベースモデルを用い、推論時における応答遅延を 0.5 秒以下という極めて低い水準に抑えることに成功した。この性能は、汎用的な GPU インフラである NVIDIA A6000 を利用してストリーミング推論を実施した際に達成されたものである。なお、本プロジェクトにおける応答遅延の定義は、音声認識（入力：音声、出力：テキスト）、音声翻訳（入力：音声、出力：テキスト）、および音声合成（入力：テキスト、出力：音声）の各処理において、それぞれの入力に対してどの程度迅速に出力を生成できるという指標で測定されている。

- **音声-to-音声の end-to-end 応答**：音声入力から音声出力までを一貫して処理可能な end-to-end モデルに関する研究はこれまで十分に確立されておらず、先行研究が非常に限られていた。そのため、安定した動作を実現するための最適なアーキテクチャの探索は困難を極めた。しかし、本プロジェクトでは広範かつ体系的なアーキテクチャ探索および多数の試行錯誤を繰り返すことによって、最終的に動作の安定性が高く実用性のあるモデル構成を特定することができ、音声入力から音声出力まで一貫して処理可能な初期モデルの開発に成功した。これにより、音声-to-音声技術のさらなる発展と応用への道筋を示した。この成果における初期のデモシステムは以下のビデオリンクから視聴する事ができる：
  https://www.youtube.com/watch?v=MK38rUBqou4



音声-to-音声の end-to-end システムの様子

- **本プロジェクトに関連する成果**：本プロジェクトに派生する成果及び今後の展望について：本プロジェクトで得られた技術成果を基盤として、Kotoba 社の独自の豊富なリソースを最大限に活用し、さらなる大規模なスケールアップと高度な技術的拡張を精力的に推進している。具体的には、リアルタイムでの音声-to-音声翻訳技術の高度化を進めており、実際のコミュニケーションにおいて自然で滑らかなインタラクションが可能なデモンストレーションを実現する段階にまで開発を進めている。また、翻訳内容の正確さだけでなく、話し手が持つ感情表現やイントネーションの微妙なニュアンスまで忠実かつ精緻に再現することを目標に掲げ、研究開発を進めている。これにより、ユーザーが音声翻訳システムを使用する際に、単なる情報の伝達を超えて、感情的および文化的な理解の深化を促すコミュニケーション支援ツールとしての

役割を担えるよう、さらなる技術的成熟度を追求している。

- 参考：音声から音声へのリアルタイム翻訳デモ：https://www.youtube.com/watch?v=7K-4bgPBPIo&ab_channel=KotobaTechnologies

現段階では日本語および英語を主要対応言語としているが、将来的にはグローバルなニーズに応えるべく、対応言語のさらなる拡充を積極的に検討している。具体的には、中国語、韓国語、アジア圏の言語などを中心に多言語化を進め、多言語間でのリアルタイムコミュニケーションを可能とする包括的なプラットフォームの構築を目指している。また、今後の技術的展望としては、本プロジェクトの研究開発プロセスにおいて得られた知見をもとに、モデルの性能を一層向上させるために必要なさまざまな技術的課題への対応に注力している。特に、リアルタイム処理のさらなる高速化やモデル推論精度の向上、推論環境やサービングシステムの最適化および大規模並列化といった要素に重点的に取り組んでいる。加えて、実世界での実用性を高めるために、ノイズの多い環境や多人数が参加する会議などの複雑な使用シナリオにおいても安定したパフォーマンスを発揮できるよう、さらなる研究開発を積極的に推し進めていく予定である。

6. 進捗状況の自己評価と今後の展望

本プロジェクトの進捗状況については、当初設定していた研究計画および目標に対して、非常に順調な達成状況を示していると自己評価している。具体的には、計画段階で掲げていた Transformer ベースの約 80 億パラメータ規模のモデルを FSDP ライブラリを用いて問題なく学習することができ、その結果、音声入力に対する応答遅延を 0.5 秒以下に抑えた end-to-end モデルの開発に成功した。これは、リアルタイム応答が求められる様々な応用分野において極めて実用的な性能を達成したことを意味しており、研究の初期段階で設定した主要な技術目標の一つを確実に達成した。

また、当初の計画において、膨大かつ多様な音声データを収集・整備するという課題があったが、実際には約 100 万時間もの大規模なデータセット（日本語 60 万時間、英語 40 万時間）の構築に成功したことも大きな成果である。この音声データセットは、公開データセットや W マイニングに加え、独自の音声合成（TTS）技術による合成データも活用して作成された。これにより、多様性と量の両面で世界的にも類を見ない規模のデータセットが構築され、今後の研究や実用化に向けたさらなる基盤整備が行われたことは特筆に値する。

技術的な詳細に踏み込むと、音声トークナイザーに関しても調査を行った。Mimi Tokenizer をはじめとする複数の音声トークナイザーを実際に検証し、それぞれの特徴を整理した。その結果、特に音声認識タスクにおいては Mimi Tokenizer が高効率かつ実用性が高いことが確認された一方、音声生成においては DAC が特定領域において高い性能を示すことも新たに発見された。このように、具体的な技術検証を通じて、実際の応用分野に応じた最適なトークナイザーの選択が可能となる知見が得られた点は、計画以上の成果と評価している。

今後の展望としては、*「5. 今年度の研究成果の詳細」*にも一部記載したように、音声-to-音声のモデルの特徴的な能力として現時点の成果をさらに発展させるために、まずは音声生成および翻訳における感情表現やイントネーションの再現性向上を目指す。人間が自然にコミュニケーションを行う際に無意識に活用している非言語的情報をモデルが正確に捉え、それを音声出力に忠実に反映できるような技術開発を進めていくことを計画している。また日本語英語だけの 2 言語出あった本プロジェクトのスコープを拡大して、多言語展開していく事も大きな目標の一つである。

※7. 研究業績はウェブ入力です