

TSUBAME 共同利用 令和6年度 産業利用 成果報告書

利用課題名 深層学習を用いたリアルタイム AI 音声合成プラットフォームの技術開発

英文: Real-time AI Voice Synthesis Platform using Deep Learning

利用課題責任者 竹内 雅樹

Masaki Takeuchi

所属 株式会社 CoeFont

CoeFont Co., Ltd.

URL <https://www.coefont.com/ja/>

邦文抄録(300 字程度)

本研究では、深層学習技術を活用したリアルタイム AI 音声合成プラットフォームの開発を行った。具体的には、長時間のゼロショット音声合成を可能にする階層型ニューラルコーデック言語モデル「HALL-E」の手法を参考にし、独自の音声合成モデルを構築した。さらに最新の AI 音声モデル、「CoeFont v3 Fuji」を開発した。CoeFont が独自に開発したであり、より人間らしさを追求した。また内閣府との意見交換会を通じて、リアルタイム AI 通訳サービス「CoeFont 通訳」の社会実装に向けた議論を行った。これらの成果により、高精度かつリアルタイムな音声合成・通訳システムの実現に寄与した。

英文抄録(100 words 程度)

In this study, we developed a real-time AI voice synthesis platform utilizing deep learning techniques. Specifically, we constructed a proprietary speech synthesis model inspired by the "HALL-E" hierarchical neural codec language model, which enables minute-long zero-shot speech synthesis. In addition, we developed the latest AI voice model, CoeFont v3 Fuji, which is proprietary to CoeFont and is designed to be more human-like. Additionally, through discussions with the Cabinet Office, we explored the societal implementation of the real-time AI interpretation service "CoeFont Interpreter." These achievements contribute to the realization of high-precision and real-time voice synthesis and interpretation systems.

Keywords: Deep learning, real-time, AI voice synthesis, zero-shot, AI interpretation

背景と目的

近年、AI 技術の進展により、音声合成や通訳システムの高度化が求められている。特に、リアルタイムで高精度な音声合成を実現することは、多言語コミュニケーションの円滑化やアクセシビリティの向上に寄与する。しかし、従来の音声合成モデルでは、長時間の音声生成やゼロショットでの多様な話者の再現が課題となっていた。

本プロジェクトでは、これらの課題を解決するため、深層学習を用いたリアルタイム AI 音声合成プラットフォームの技術開発を行い、高精度かつリアルタイムな音声合成・通訳システムの実現を目指した。

概要

本研究では、以下の取り組みを行った：

1. 階層型ニューラルコーデック言語モデル「HALL-E」による音声合成モデルの構築

HALL-E は、長時間のゼロショット音声合成を可能にするモデルであり、マルチレゾリューション再量子化(MReQ)と呼ばれる手法を用いて、ニューラルオーディオコーデック(NAC)のフレームレートを低減する。これにより、1 分以上の音声を安定して生成できる[1]。

2. 最新の音声合成モデルおよび通訳モデルの開発

深層学習モデルを用いて、「CoeFont v3 Fuji」音声合成モデルを開発した。主な特長は人間の発声特性を解析し、抑揚やリズムを忠実に再現し、AI 音声でありながら、話者の感情やニュアンスをよりリアルに再現し現時点での日本語での AI 音声において、最高水準の音声表現が可能となった。これにより、自然さと感情表現で大幅な改善がなされた[2]。また、翻訳処理を統合したリアルタイム通訳機能「CoeFont 通訳」も開発した。TSUBAME を活用した高速な学習・推論環境の構築により、実用化に向けた安定稼働を実現した[3]。

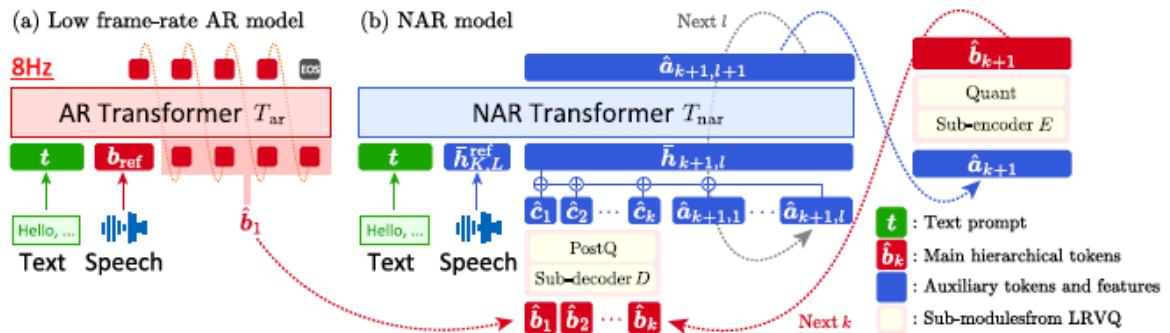


図 1. HALL-E のアーキテクチャ

結果および考察

本研究の結果、以下の成果を得た。

• 高精度な長時間音声合成の実現

HALL-E は従来の音声合成モデルに比べて以下のような顕著な利点を有している。

- 1 分以上の長時間音声をゼロショットで高精度に生成することが可能
- フレームレートを効果的に圧縮し、音質を維持しながらトークン長を最大 96%削減。生成速度と効率を飛躍的に向上。
- 音声を複数レベルの符号列(トークン)に変換。粗い特徴から詳細な特徴へと段階的に復元するため、長時間生成でも破綻しにくい。
- 話者向けに再構成したニューラルオーディオコーデック(NAC)を訓練し、HALL-E に類似した階層的トークン生成と、文脈保持に優れた言語モデルを統合し、CoeFont が保有する大規模日本語音声データを用いたチューニングした結果、従来モデルと比較して WER や MOS 等の指標において性能向上が確認された。図 1 に、HALL-E のアーキテクチャを示す。また実際にモデルから生成された音声のデモは[4]より聞くことができる。

考察として、階層的トークン設計により、従来破綻しがちだった長時間音声の自然な合成が可能となり、ユーザーによる連続的な対話応答やナレーション用途での利用が現実的になった。話者固有の音響特徴(声色、話し方)を数秒の音声から抽出し、それを使って任意テキストの音声合成を行えるため、個別の学習コストを削減しながら多様なユーザー対応が可能になった。

MReQ ベースのトークン圧縮により、生成レイテンシは大幅に短縮され、リアルタイム用途(例:会議同時通訳、対話システム)でも実用可能な応答速度を達成した。

また内閣府との意見交換会を通じて、「CoeFont 通訳」の実用性を示し、社会実装に向けた具体的な課題や展望を共有することができた。これらの成果は、リアルタイムで高品質な音声合成・通訳システムの実現に向けた重要な一歩となった。

まとめ、今後の課題

本研究では、深層学習を用いたリアルタイム AI 音声合成プラットフォームの技術開発を行い、高精度な長時間音声合成やリアルタイム通訳システムの実現に向けた成果を得た。

まず CoeFont v3 Fuji およびリアルタイム通訳機能を開発・実証し、商用展開に向けた技術的基盤を確立した。今後は、より多様な言語への対応、感情と音韻の統合表現、倫理的ガイドライン整備を進め、医療・教育・放送・障害者支援等の分野への展開を図る。TSUBAMEとの連携を継続し、大規模モデル学習と評価も並行して進める予定である。

今後の課題として、多言語対応が挙げられる。現状のシステムは、日本語における音声認識・合成・翻訳に高い精度を示しているが、今後の実用展開を見据えると、英語を含めた多様な言語への対応が求められる。特に、東南アジア諸国や中南米、アフリカ諸国など、英語圏以外の地域におけるグローバルな利用を想定した場合、現地言語の認識・翻訳性能の向上が必要不可欠である。

具体的には、以下の点が今後の技術的課題となる：

- 言語ごとの発話特徴や文法構造に最適化した

音響モデル・テキスト処理モデルの構築

- 話者数・アクセントの多様性を含む音声データセットの収集と精緻なアノテーション
- 翻訳先の言語文化・文脈を反映した自然な言い換えや口語表現への対応

また、将来的には、ユーザーが任意の言語ペアを自由に選択できる汎用的なリアルタイム音声翻訳フレームワークの構築を目指す。これにより、ビジネス・観光・教育・国際会議など、あらゆるシーンで活用可能な多言語音声プラットフォームの実現が期待される。

参考文献

[1] Yuto Nishimura, Takumi Hirose, Masanari Ohi, Hideki Nakayama, Nakamasa Inoue. “HALL-E: Hierarchical Neural Codec Language Model for Minute-Long Zero-Shot Text-to-Speech Synthesis”, International Conference on Learning Representations (ICLR), 2025.

<https://arxiv.org/abs/2410.04380>

[2] CoeFont、AI 音声の新時代へ！新モデル「CoeFont v3 Fuji」実装で“人らしさ”がさらに進化
<https://prtimes.jp/main/html/rd/p/000000099.000078329.html>

[3] CoeFont、リアルタイム AI 通訳「CoeFont 通訳」の活用等に関して内閣府と意見交換会を開催
<https://prtimes.jp/main/html/rd/p/000000096.000078329.html>

[4] HALL-E: Hierarchical Neural Codec Language Model for Minute-Long Zero-Shot Text-to-Speech Synthesis
https://yutonishimura-v2.github.io/HALL-E_DEMO.