TSUBAME 共同利用　令和7年度　学術利用　成果報告書

利用課題名　大規模言語モデル用いた日本語音声合成・対話システムの開発
英文　Developing Large Language Model-based Japanese Speech Synthesis and Dialogue Systems

利用課題責任者
HUANG Wen-Chin

所属　名古屋大学情報学研究科
https://www.i.nagoya-u.ac.jp/graduate-school-of-informatics/

邦文抄録（300 字程度）

本課題では，大規模言語モデル（LLM）を活用した日本語音声合成・対話システム技術の開発を目的として，日本語音声合成および EL 音声変換の二つの課題に取り組んだ。実験結果として，現時点のシステム性能は十分でなく，特に明瞭度の改善が課題として残った。一方で，一部の合成音声には自然な特徴も確認され，LLM ベース音声合成・対話システム技術の日本語応用に一定の可能性が示された。また、日本語における LLM ベース音声合成モデルの性能向上には，大規模な学習データの収集が重要であることが分かった。

英文抄録（100 words 程度）

The goal of this project is to develop large language model (LLM)-based Japanese speech synthesis and dialogue systems. We experimented with two task: text-to-speech, and electrolaryngeal speech voice conversion. Experimental results demonstrated that the performance of our current systems is not yet satisfactory but is showing potential. Our results also highlighted the importance of collecting large-scale training datasets.

Keywords: speech synthesis, speech dialogue system, large language model

背景と目的

近年，大規模言語モデル（large language model; LLM）を活用した音声合成（text-to-speech; TTS）・対話技術が注目を集めている [1]。LLM は，テキストや音声に含まれる文脈情報を高精度にモデル化できることから，従来の TTS 技術では困難であった，発話内容や対話文脈に応じた自然な音声生成への応用が期待されている。

一方で，既存の研究の多くは，英語や中国語など，音声・テキストデータが比較的豊富に存在する言語を対象としている。そのため，日本語のように利用可能な音声データが限られている言語において，LLM を用いた TTS・対話技術がどの程度有効に機能するかについては，十分に検証されていない。また，日本語は独自の文字体系を持つだけでなく，ピッチアクセントやイントネーション，抑揚などの制御が音声の自然性に大きく影響する。このような言語的・音声的特徴を適切に捉え，文脈に応じた発話スタイルを生成することは，現在の日本語 TTS・対話技術における重要な課題である。

本課題では，LLM を活用することで，日本語 TTS・対話におけるデータ不足や日本語特有の言語特性に起因する課題の解決を目指した。具体的には，日本語データを用いて LLM を学習し，その高い文脈モデリング能力および特徴制御能力を活かすことで，文脈に応じた音声特徴の制御を行い，自然で文脈に適応した応答が可能な対話音声生成手法の実現に取り組んだ。

概要

本課題では，LLM を活用した日本語音声合成技術に関する研究を実施した。特に，日本語のような音声データ資源が限られた環境において，大規模モデルが有効に機能するかを検証するとともに，音声合成および音声変換への応用可能性を調査した。

1. 日本語音声合成における LLM ベースモデルの有効性検証

まず，日本語TTSを対象として，LLMベースのTTSモデルである VALL-E の学習および評価を行った。本研究グループでは，これまでに日本語TTS研究を支援

	学習データ	自然性指標↑	明瞭度指標↓	話者類似度指標↑
FastSpeech2	JVS（約 16 時間）	4.49	3.0	0.709
VALL-E	CSJ（約 515 時間）	2.61	64.8	0.577

表 1: 従来型音声合成モデル FastSpeech2 と LLM ベース音声合成モデル VALL-E の客観評価結果

	自然性指標↑	明瞭度指標↓
正解音声	3.92	12.0
EL-Moshi	3.82	50.9

表 2：正解音声と EL-Moshi による生成音声の客観評価結果

するオープンソースツールキット JATTS [2] を開発しており，従来は FastSpeech2 [3], VITS [4], Matcha-TTS [5] などのモデルを対象としていた。また，学習には JSUT [6]（約 10 時間），JVS [7]（約 16 時間），HiFi-Captain [8]（約 20 時間）といった比較的小規模な日本語音声データセットを用いていた。本課題では，JATTS を拡張し，大規模日本語話し言葉コーパスである CSJ [9]（約 515 時間）を用いて VALL-E [10]の学習を実施した。VALL-E は，元々約 60,000 時間規模の音声データで学習された LLM ベース TTS モデルであるため，日本語において利用可能な 515 時間程度のデータ量でも十分な性能が得られるかを検証した。

2. 日本語 EL 音声変換への LLM ベース音声対話モデルの応用

次に，音声変換技術の応用として，EL（Electrolaryngeal）音声変換システムの開発に取り組んだ。EL 音声とは，喉頭全摘出手術を受けた患者が電気式人工喉頭（Electrolarynx）を用いて発声した音声であり，発話が可能となる一方で，機械的で不自然な音質となることが課題である。本研究では，EL 音声を自然な音声へ変換することで，患者の QOL（Quality of Life）向上に貢献することを目指した。具体的には，リアルタイム音声対話システムとして知られる LLM ベースの音声モデル Moshi [11] に着目し，これを用いた日本語 EL 音声変換システムの構築を試みた。Moshi は元々約 700 万時間規模の英語音声データを用いて学習されているが，日本語において同規模のデータ収集は現実的ではない。そこで，本課題では約 2,000 時間の日本語音声データを用いてモデルの追加学習を行い，日本語環境における有効性を検証した。

結果および考察

1. 日本語音声合成における LLM ベースモデルの有効性検証

表 1 に日本語 TTS モデルの客観評価結果を示す。比較対象として，従来型の TTS モデルである FastSpeech2 を用いた。FastSpeech2 は約 16 時間の JVS データセットを用いて学習したモデルである。一方，本課題で構築した VALL-E は，約 515 時間の CSJ を用いて学習した LLM ベースの TTS モデルである。

客観評価の結果，VALL-E は FastSpeech2 と比較して，自然性指標，明瞭度指標，話者類似度指標のいずれにおいても低い性能を示した。特に，明瞭度指標の劣化が大きく，生成音声の発話内容を安定して再現することが課題として確認された。これは，LLM ベース TTS モデルが高い表現能力を持つ一方で，十分な性能を発揮するためには大規模な学習データを必要とすることを示唆している。実際，元の VALL-E は約 60,000 時間規模の音声データを用いて学習されており，本課題で用いた 515 時間の日本語音声データは，それと比較すると大幅に小さい。

一方で，合成音声を確認したところ，一部のサンプルでは自然な韻律や音声特徴が確認され，LLM ベース TTS モデルとしての可能性も見られた。したがって，現段階では既存の従来型モデルを上回る性能には至らなかったものの，日本語における LLM ベース TTS の実現可能性を検証する上で有用な知見が得られた。

2. 日本語 EL 音声変換への LLM ベース音声対話モ

デルの応用

表 2 に，LLM ベース音声対話モデル Moshi を用いた日本語 EL 音声変換の評価結果を示す。本課題では，EL 音声データを用いて追加学習を行った Moshi モデルを EL-Moshi と呼ぶ。比較対象として，正解音声の評価結果も併せて示した。評価の結果，EL-Moshi は自然性指標において正解音声に近い値を示した。一方で，明瞭度指標については正解音声より大きく低下しており，発話内容の明瞭な再現には依然として課題が残ることが分かった。これは，表 1 で示した VALL-E の結果と一致し，LLM ベースの音声生成モデルでは，自然な音声特徴の生成に一定の可能性がある一方で，明瞭で正確な発話内容の生成にはより大規模なデータが必要であることを示唆している。

まとめ、今後の課題

本課題では，LLM を活用した日本語音声合成技術について，日本語 TTS および日本語 EL 音声変換の二つの観点から検討を行った。VALL-E を用いた日本語音声合成では，従来型モデルと比較して客観評価結果は十分ではなかったものの，一部の生成音声には自然な特徴も確認された。また，Moshi を用いた EL 音声変換では，自然性には一定の可能性が見られた一方で，明瞭度の改善が課題として残った。これらの結果から，LLM ベース音声合成モデルを日本語へ適用するには，特にデータ量の確保と明瞭度の改善が重要であることが分かった。今後は，VALL-E より新しい LLM ベース TTS 手法を導入し，日本語 TTS における有効性をさらに検証する。また，近年主流となっている Web データの活用により，日本語音声データの規模を拡大し，モデルの自然性および明瞭度の向上を目指す。

参考文献

[1] Arora, Siddhant, et al. "On The Landscape of Spoken Language Models: A Comprehensive Survey." Transactions on Machine Learning Research, 2025.

[2] W.-C. Huang, L.P. Violeta, T. Toda, "JATTS: a comparison-oriented Japanese text-to-speech open-sourced toolkit," 信学技報, Vol. 125, No. 74, SP2025-22, pp. 119-124, June 2025.

[3] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.Y. Liu, “FastSpeech 2: Fast and High-Quality End-to-End Text to Speech,” Proc. ICLR, 2021.

[4] J. Kim, J. Kong, and J. Son, “Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech,” Proc. ICML, pp.5530–5540, 2021.

[5] S. Mehta, R. Tu, J. Beskow, E. Sz´ekely, and G.E. Henter, “Matcha-TTS: A fast TTS architecture with conditional flow matching,” Proc. ICASSP, 2024.

[6] R. Sonobe, S. Takamichi, and H. Saruwatari, “JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis,” arXiv preprint arXiv:1711.00354, 2017.

[7] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, “JVS corpus: free Japanese multi-speaker voice corpus,” arXiv preprint arXiv:1908.06248, 2019.

[8] T. Okamoto, Y. Shiga, and H. Kawai, “Hi-Fi-CAPTAIN: High-fidelity and high-capacity conversational speech synthesis corpus developed by NICT.” https://ast-astrec.nict.go.jp/en/release/hi-fi-captain/, 2023.

[9] K. Maekawa, “Corpus of spontaneous Japanese: its design and evaluation,” Proc. ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003.

[10] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, et al., “Neural codec language models are zero-shot text to speech synthesizers,” arXiv preprint arXiv:2301.02111, 2023.

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