音学シンポジウム2026

11

キャラクター画像を用いた日本語テキスト音声合成の声質制御

見た目は口ほどにものをいう？

◎ 周 廷叡[1], 柳井 啓司[1]

[1]電気通信大学 I類 メディア情報学プログラム

背景

- Text-to-Speech (TTS) で声質の制御を行う既存の手法は、**音声参照**や**テキストキャプション**による制御が主流
- 一方で、キャラクターのイメージが既に定まっている場合は**見た目と声の雰囲気を合わせるのが困難**

目的

キャラクターの画像を入力し、その**キャラクターのイメージに沿う声**で TTS を行う手法 **Character Voice Control (CVC)** を提案する。

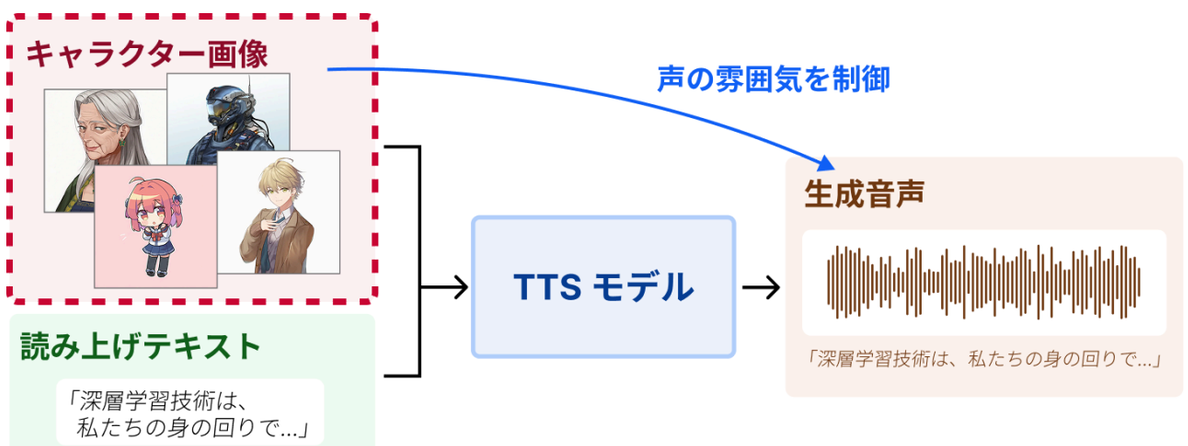


図1: 全体の処理の流れ

提案手法: Character Voice Control (CVC)

事前学習済み TTS モデルとして **Irodori-TTS**[1] を採用した。参照画像を追加条件として入力するため、提案手法では新たに**事前学習済み画像エンコーダーと MLP** を導入した。
学習・推論時はこれらのモジュールにキャラクター画像を入力する。 実験では、追加した **MLP 層のみを学習対象**とし、他のモジュールはすべて凍結した。 画像エンコーダーは二種類検討した。

- **CVC-Tagger**: wd-tagger-v3-vit[2]
- **CVC-SigLIP**: SigLIP-v2-B/16-512[3]

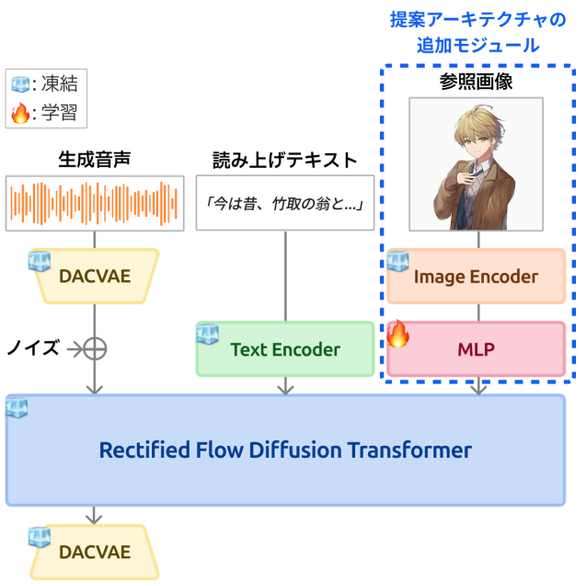


図2: 提案手法のアーキテクチャ

図3 は DiT ブロック内で行われる Joint Attention 処理を示している。 参照音声特徴を Key, Value に変換する**事前学習済みの重みを再利用**して、参照画像特徴を Key, Value に変換する。これらの Key, Value は音声潜在変数やテキスト特徴の Key, Value と concat された後、Self Attention によって in-context で条件付けされる。

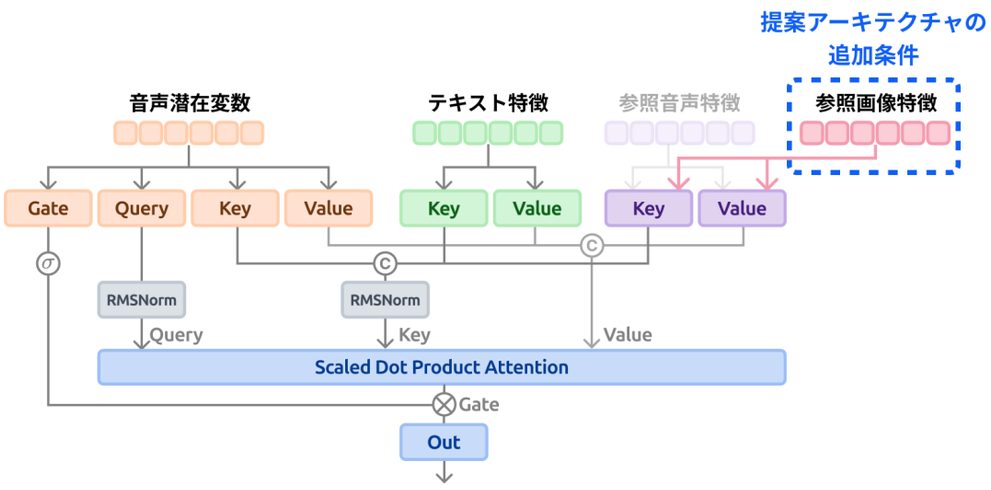


図3: Rectified Diffusion Transformer ブロック内の Joint Attention 処理

データセット

学習には、**キャラクター画像・音声・書き起こしテキストの三つ組ペアデータセットが必要**となる。 本実験では、インターネット上から**独自に収集**してデータセットを構築した。 まず、14,000 件の画像-音声ペアを収集した。 それらのうち、手作業や VAD によって、画像から音声が想像しづらいものや人の声でないものを除外した。 その後、Whisper Large v3 [4] を用いて音声を書き起こした。後処理後、書き起こしが1文字以上残ったものを選定し、**最終的に 6,768 ペア**が選定された。

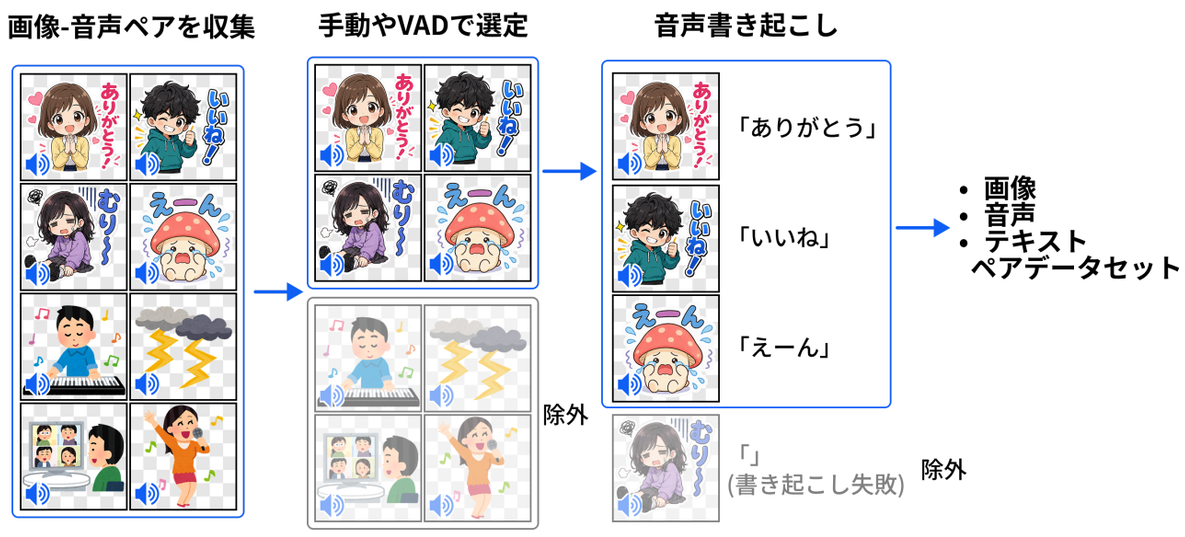


図4: データセットフィルタリングの流れ

実験・結果

学習は、CVC-Tagger、CVC-SigLIP ともに、1 台の NVIDIA RTX 4070 Ti Super を用いて約 5 時間かけて行われた。
評価では、5 段階 Mean Opinion Score (MOS) で自然性と親和性を評価した。 音声の自然性を客観的に評価するためのシステム UTMOSv2 を用いた **自然性 MOS** では、**提案手法はベースラインと同程度**の評価となった(表1)。
人手によるキャラクター画像と生成音声の**親和性 MOS** では、**提案手法がベースラインを上回った**(表2)。

表1: UTMOSv2 自然性 MOS

TTS condition	UTMOSv2 ↑
No reference	$\underline{2.593 \pm 0.043}$
Voice Design	2.482 ± 0.044
CVC-Tagger (Ours)	$\mathbf{2.651 \pm 0.039}$
CVC-SigLIP (Ours)	2.556 ± 0.037

表2: 人手による画像-音声親和性 MOS

TTS condition	Affinity MOS ↑
No reference	2.514 ± 0.507
Voice Design	3.186 ± 0.636
CVC-Tagger (Ours)	$\underline{4.029 \pm 0.279}$
CVC-SigLIP (Ours)	$\mathbf{4.400 \pm 0.200}$

マルチモーダル LLM (MLLM) を用いた評価も検討した。 図5 は人手と MLLM (Gemini 3.1 Pro, Gemini 3.1 Flash, Qwen3 Omni Thinking) による親和性 MOS の散布図を示している。 Gemini 3.1 Pro は比較的人間の評価と相関係数が高い(r = 0.628)が、人間の評価者を代替できるとは言えない結果となった。

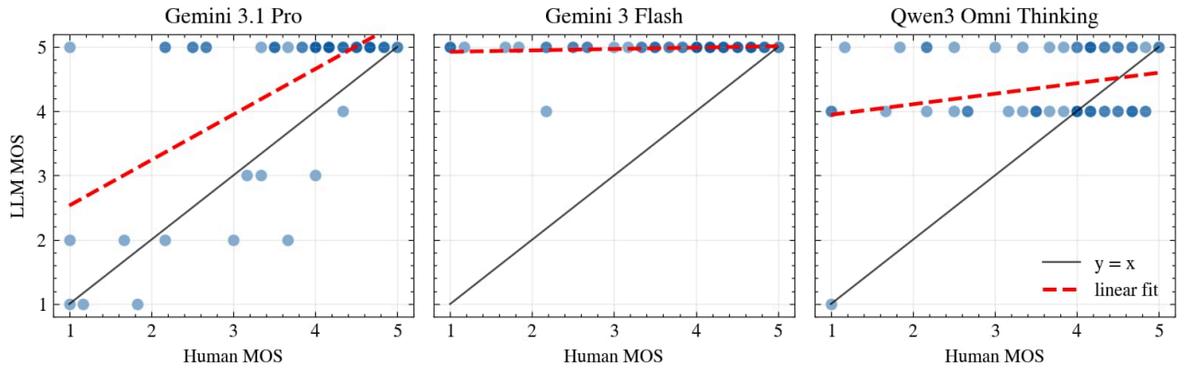


図5: 人手と MLLM による親和性 MOS の散布図

結論

事前学習済みの TTS モデルと画像エンコーダーを用いた、**キャラクター画像を参照**することで**「キャラクターらしい声」で TTS を行う手法を提案**し、人手による親和性 MOS ではベースラインよりも高い評価となった。 MLLM を用いた評価は人間を代替するには至らず、今後の課題として画像-音声間の親和性評価指標が求められる。

参考文献

1. Chihiro Arata. Irodori-TTS: A flow matching-based text-to-speech model with emoji-driven style control. GitHub, https://github.com/Aratako/Irodori-TTS, 2026.
2. SmilingWolf. WD ViT Tagger v3. Hugging Face, https://huggingface.co/SmilingWolf/wd-vit-tagger-v3, 2024.
3. Michael Tschannen, Alexey Gritsenko, Xiao Wang, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786, 2025.
4. Alec Radford, Jong Wook Kim, Tao Xu, et al. Robust speech recognition via large-scale weak supervision. In Proc. ICML, pages 28492–28518, 2023.