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

周 廷叡[1,a)]　柳井啓司[1]

概要：本研究では，アニメやゲームなどのキャラクター画像を入力として声質を制御する日本語テキスト音声合成手法を提案する．従来のテキスト音声合成では，音声を参照して声質を制御する Voice Cloning やテキストキャプションで制御する Voice Design が用いられてきたが，キャラクターの外見から直接声質を推定・制御する手法は十分に確立されていない．提案手法では，フローマッチングに基づく事前学習済み音声合成モデルに画像エンコーダーを統合し，キャラクター画像から声質を条件付ける枠組みを構築した．提案手法はベースラインと比較して UTMOSv2 自然性評価を維持しつつ，人間による親和性 MOS 評価で Voice Design を上回った．また，マルチモーダル LLM を用いた評価も検討し，その限界を明らかにした．

キーワード：テキスト音声合成，声質制御，キャラクター画像

Character Image-Conditioned Speaker Style Control for Japanese Text-to-Speech

Abstract: In this study, we propose a Japanese text-to-speech (TTS) method that controls voice style using character images from anime and games as input. In conventional TTS, voice style is typically controlled through Voice Cloning, which uses reference audio, or Voice Design, which uses text captions. However, methods for directly estimating and controlling voice quality based on a character's visual appearance have not been sufficiently established. The proposed method integrates an image encoder into a pre-trained flow matching-based TTS model, constructing a framework that conditions speaker style on character images. Compared to the baseline, our method maintained speech naturalness in UTMOSv2 evaluations while outperforming Voice Design in human Mean Opinion Score evaluations for character-voice affinity. Additionally, we investigated the use of Multimodal Large Language Models for evaluation and identified their limitations.

Keywords: Text-to-Speech, Voice Style Control, Character Image

1. はじめに

アニメや映画，ゲームなどのメディア作品では，キャラクターの容姿や性格に対して「それらしい声」が存在するという強い期待が視聴者・ユーザーに共有されている．そのような期待を背景にして，近年の Seedance 2.0 [1] などの動画生成モデルでは，音声が映像の内容と合致するような動画の生成が可能となってきている．キャラクターの外見・振る舞い・発話内容が同時に生成される場面では，声質もその視覚的な印象と整合している必要がある．また，昨今では誰もが独自の AI エージェントや AITuber のようにキャラクターに声を与える場面が一般化しつつあり，そのキャラクターのイメージに合致した独自の声を付与したいという需要は今後さらに高まると考えられる．

しかし，現在のテキスト音声合成（TTS）技術においては，キャラクターのイメージを基にして，自然と受け入れられるような声質へと直感的に制御する手法は十分に確立されていない．今日では TTS 技術はナレーションや音声対話型インターフェースへの利用が進んでいるものの，その多くは，事前に特定の話者の音声で学習されたモデルを用いる方法や，音声から声質を参照してその声質で音声合成を行う手法が主流である．

それに対して近年，単一モデルで多様な話者や声質の音

[1] 電気通信大学
The University of Electro-Communications
[a)] zhou-t@mm.inf.uec.ac.jp

© 1959 Information Processing Society of Japan

声を合成可能な技術が登場している．Qwen3-TTS [2] や OmniVoice [3]，VoxCPM [4] は多言語データセットを用いて大規模に学習され，単一のモデルを用いてさまざまな言語，声質で音声を合成することができる．これらの手法は，音声を参照して声質を制御する方法 Voice Cloning だけでなく，参照すべき音声がない状態から独自の声質で音声合成を行う方法として，テキストキャプションで声質を指定する手段 Voice Design が提供されている．

これらの手法の登場によって多様な声を生成できるようになったものの，独自の声質で音声を合成したい場合の制御性にはまだ課題がある．Voice Cloning を用いるには参照音声が存在しなければならない．そのため、何らかの方法で参照となる音声を用意する必要があり，独自の声質を構築したい場合は適用が容易ではない．また，Voice Design では自然言語を用いて声質を表現する必要があるが，言語化できない細かいニュアンスを表現することは難しい．したがって，声を割り当てたいキャラクターやペルソナが存在する場合，そのイメージにあった声質での音声合成は容易ではない．

自然言語による制御が困難な場合，画像によって制御するという選択肢が存在する．テキストから画像を生成する技術は Stable Diffusion XL [5] をはじめとして，テキストキャプションを記述することで様々なコンテンツを指定できる．一方で，アートスタイルや細かい構図の指示など，言語化しづらい概念は制御が難しいという課題があった．そのような課題に対して，ControlNet [6] や IP-Adapter [7] などの手法は，画像を追加の入力として用いることで直感的な制御を実現している．

そこで本研究では，音声を割り当てたいキャラクターが存在する場合にフォーカスして，画像を用いた声質制御を行う TTS 手法 Character Voice Control (CVC) を提案する．CVC では，日本語で事前学習された Irodori-TTS [8] をベースとして，キャラクター画像を入力に追加することで，そのキャラクターのイメージにあった声質での音声合成を目指す．

本研究の主な貢献は以下の三点である．

- キャラクター画像，音声，発話テキストを対応付けた三つ組データセットを構築し，画像に基づく声質制御の学習および評価に利用した．
- 事前学習済み TTS モデルに画像特徴量を統合することで，キャラクター画像を条件とした日本語音声合成を実現する枠組みを構築した．
- 人手評価に加えてマルチモーダル LLM を用いた評価を行い，キャラクター画像と合成音声の適合性評価における両者の関係を分析した．

2. 関連研究

2.1 顔画像を条件とした音声合成

Face2Speech

Face2Speech [9] は実写の顔画像を入力として，その顔画像に対応する話者の声を合成する手法である．顔画像と対応づいた音声のペアについて，それぞれ顔画像エンコーダーと音声エンコーダーに入力し，出力される特徴量を GE2E[10] 損失で最適化する．これによって学習された音声エンコーダーを，複数話者 TTS が可能なモデルに接続して学習後，音声エンコーダーを顔画像エンコーダーに置き換えることで，顔画像を条件とした音声合成を実現している．

本研究との相違点として，Face2Speech は実写の顔画像に限定されているのに対し，本研究では顔立ち・表情・服装・姿勢・背景や装飾的な視覚表現など，キャラクターイラスト全体から受ける視覚印象を条件として声質を制御する点が挙げられる．また，Face2Speech では画像エンコーダーの出力をあらかじめ対照学習によって話者特徴量空間にマッピングするのに対し，本研究ではエンドツーエンドの学習を通じて画像特徴を音声特徴空間にマッピングする点も異なる．

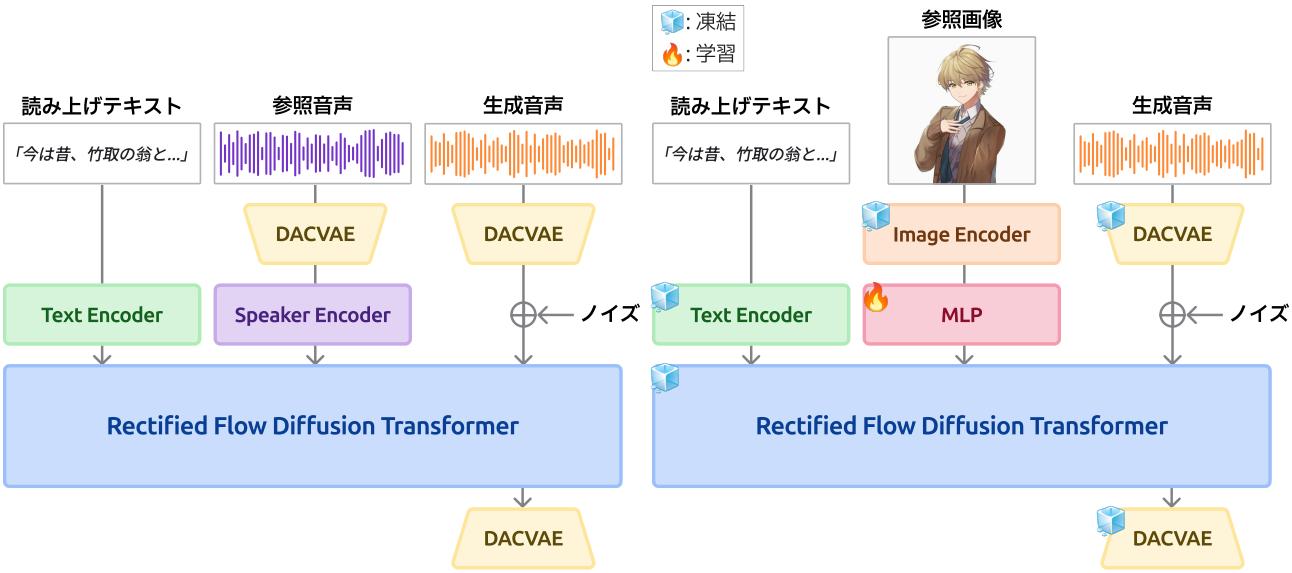
F2S-SBV2

菊池ら [11], [12] は，アニメ調キャラクター顔画像にフォーカスしたテキスト音声合成を提案している．F2S-SBV2 [11] はアニメ調キャラクター顔画像を入力に追加した Style-Bert-VITS2 (SBV2)[*1] である．既存の SBV2 はゼロショットの複数話者 TTS は行えない．そこで，F2S-SBV2 では一度事前学習済み音声エンコーダーを入力に追加することで複数話者での生成を可能にした Multi-Speaker SBV2 (MS-SBV2) を構築した．その後，顔画像で事前学習された画像エンコーダーの出力特徴量を，音声エンコーダーの特徴量空間へ写像するように MLP を学習している．そうして得られた顔画像エンコーダーと MLP の出力を MS-SBV2 の入力に追加することで，顔画像を条件とした音声合成を実現している．また，菊池ら [13] は顔画像の特徴量だけでなく，話者の離散的な年齢情報を入力に追加することで，出力を改善する手法も提案している．

F2S-SBV2 の画像ドメインはアニメイラストではあるものの，Face2Speech と同様に顔領域にフォーカスしており，服装や姿勢などの情報は考慮されていない．これは，服装から想起されるステレオタイプの情報を失う可能性があるほか，顔や頭部を持たないキャラクターには適用できないという問題がある．対して，本研究では顔領域に限定しないことで，より広い範囲のキャラクター画像を条件にすることが可能である．

[*1] https://github.com/litagin02/Style-Bert-VITS2

© 1959 Information Processing Society of Japan



(a) Irodori-TTS 音声参照時のアーキテクチャ　　(b) 提案手法 CVC のアーキテクチャ

図 1: Irodori-TTS と提案手法のアーキテクチャ図

2.2 Irodori-TTS

Irodori-TTS [8] は Echo-TTS [14] の手法をベースに構築された Rectified Flow Diffusion Transformer (RFDiT) [15], [16] モデルであり，Irodori-TTS-500M-v2 [17] は約 5 万時間の日本語音声データセットを用いて学習された日本語 TTS モデルである．

図 1a に Irodori-TTS のアーキテクチャ図を示す．Irodori-TTS では，読み上げるテキストを llm-jp/llm-jp-3-150m [*2] で初期化されたテキストエンコーダーでエンコードし，それを条件として音声を生成する．任意で参照音声を DACVAE [18], [19] でエンコードし，話者エンコーダーでエンコードした特徴量を条件に追加することで，声質を制御した生成を行うことが可能である．また，Irodori-TTS-500M-v2-VoiceDesign [20] では，DACVAE [19] と話者エンコーダーの代わりに，テキストキャプションエンコーダーの特徴量を同様にして条件に指定することで，テキストキャプションによる声質制御を可能にしている．

Flow Matching

Irodori-TTS [8] は最適輸送に基づく Flow Matching モデルであり，モデル $v_\theta(x_t, t, y_{\text{text}})$ は，正規分布からのサンプル $x_0 \sim \mathcal{N}(0, I)$ と実データ分布からのサンプル $x_1 \sim p_{\text{data}}(x)$ の時刻 $t \in [0, 1]$ におけるノイジーサンプル $x_t = tx_1 + (1-t)x_0$ を入力とし，読み上げテキスト y_{text} を条件にして，速度場 $v_t = x_1 - x_0$ を予測するように学習する．このとき，損失関数 $\mathcal{L}$ は以下のように表される．

$$\mathcal{L} = \mathbb{E}_{t,x_0,x_1} \| v_t - v_\theta(x_t, t, y_{\text{text}}) \|_2^2 \qquad (1)$$

また，参照音声を用いる場合の損失関数 $\mathcal{L}_{\text{ref}}$ は，参照音声をエンコードした特徴量 y_{ref} を更なる条件として加えて以下のように表される．

$$\mathcal{L}_{\text{ref}} = \mathbb{E}_{t,x_0,x_1} \| v_t - v_\theta(x_t, t, y_{\text{text}}, y_{\text{ref}}) \|_2^2 \qquad (2)$$

2.3 AFCVF データセット

菊池ら [12] は F2S-SBV2 [11], [12] の学習のためにデータセット Anime-Face-Character-Voice-Face (AFCVF) を構築した．コンピュータゲームから抽出した音声とそれぞれに対応するキャラクターのイラスト画像のペアを収集した．さらに，直接対応しないが類似する顔パーツが含まれるキャラクターのイラスト画像をペアとみなし，データセットを作成した．

しかしながら，キャラクターの容姿が似ていても，細部の違いからイメージされる声質が大きく異なる可能性があるため，本来の声質と画像の対応が曖昧になるという課題がある．例えば，同じ髪色のキャラクターであっても髪型が異なると，連想するキャラクター性のステレオタイプが変わる可能性が考えられる．本研究で用いるデータセットはこのような曖昧な対応付けを避けるため，本来の画像音声ペア以外と対応させるような拡張は行わなかった．

3. 提案手法

3.1 データセット

画像を参照したテキスト音声合成を行うには，画像・音声・テキストのペアデータセットが必要となる．一方でそのようなデータセットは一般に公開されていないため，独自に構築した．

まず，インターネット上から画像と音声が直接結びつい

[*2] https://huggingface.co/llm-jp/llm-jp-3-150m

© 1959 Information Processing Society of Japan

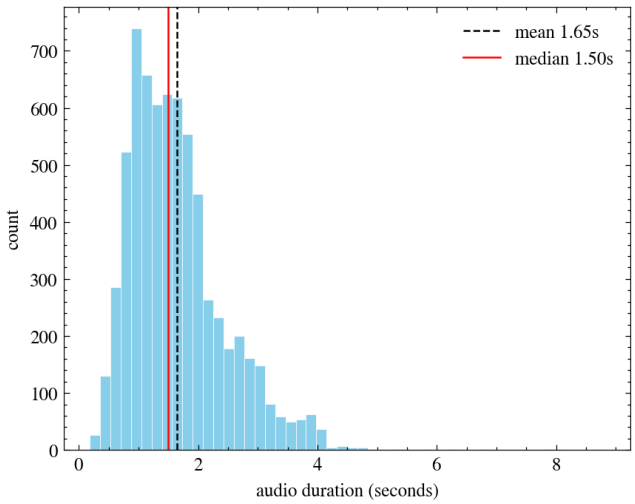


図 2: データセットの音声長分布

た 14000 ペアを収集した．このペアにはキャラクターが映らない画像や声質がイメージしづらい画像も含まれるため，画像を基に手動でフィルタリングし 7160 ペアを選定した．また，ペアの大部分はキャラクターのボイスが対応するが，声が入っていないものや BGM などのノイズが入っているケースが存在するため，Silero VAD [21] を用いて音声区間検出を行い，人の声である確率が 20% 以上のデータを自動的に採用した．一方，確率が 20% 未満のデータについては手動で確認を行い，実際に人の声が含まれていないものをデータセットから除外した．これによって 6859 ペアが選定された．

この段階では読み上げテキストが存在しないため，音声から書き起こしを行った．音声の書き起こしには，高速な推論が可能な faster-whisper [22] ライブラリを介して，OpenAI Whisper large-v3 [23] モデルを利用した．その後，書き起こしにおける同じ文字の異常な連続や，無音区間に対するハルシネーションをルールベースで取り除いた．最終的に書き起こしテキストが 1 文字以上存在する 6768 ペアを選定した．また，これらの音声に対して librosa [24] を用いて，音声の終端の無音区間をトリミングした．このようにして得られた画像・音声・テキストの 6768 ペアが，提案手法の学習に用いられた．このうち，0.5% を学習中の検証セットとして分割し，残りを学習セットとした．

得られたデータセットの音声長分布を **図 2** に示す．このデータセットでは，平均音声長が 1.65 秒，中央値が 1.50 秒，最大値が 8.81 秒であり，合計約 3 時間の音声データとなった．

このデータセットには，一つの音声で二人以上の話者が同時に発声しているものや，一つの画像に二人以上のキャラクターが写っているものも含まれた．画像に含まれる複数のキャラクターと音声内の声が適切に対応が取れる場合は，有効なデータとしてデータセットから取り除かず，学習の対象とした．

3.2 アーキテクチャ

図 1b に提案手法のアーキテクチャ図を示す．本手法では，音声参照を用いる Irodori-TTS [8] のアーキテクチャをベースに，DACVAE [19] と話者エンコーダーを，事前学習済み画像エンコーダーと MLP で置き換えた．学習時はその MLP のみを学習対象とし，画像エンコーダーや DACVAE [19]，RFDiT などの他のモジュールは全て凍結した．実験では二種類の画像エンコーダーを比較検討した．

CVC-Tagger

CVC-Tagger では wd-vit-tagger-v3 [25] を画像エンコーダーに用いた．wd-vit-tagger-v3 はイラスト画像のみを対象として，画像の内容をラベリングするタスクで事前学習されたモデルである．今回の実験では主にイラスト画像を用いて学習であるため，イラストドメインに特化した画像エンコーダーを用いることでよりイメージに沿った声質制御が可能になることを期待した．wd-vit-tagger-v3 は 12 層から成る，パッチサイズ 16 の Base サイズ ViT [26] モデルであり，448×448 の画像を入力として受け取って 768 次元の特徴量を出力する．

エンコーダーの出力を下流タスクに利用する際，最終層の出力よりも中間層の出力を用いると性能が向上することが知られている．[5], [27] そこで本実験では，最後から 5 番目の Transformer ブロックの出力を後続の MLP への入力とした．このとき，プーリング処理は行わずパッチシーケンスをそのまま MLP への入力とするため，MLP への入力サイズは パッチ数 $28 \times 28 = 784$ と出力次元 768 から $(784, 768)$ となる．

CVC-SigLIP

CVC-SigLIP では SigLIP2-B/16-512 [28], [29] を画像エンコーダーに用いた．SigLIP2-B/16-512 は画像とテキストの対照学習に基づいて事前学習されたモデルである．12 層から成るパッチサイズ 16 の Base サイズ ViT モデルで，512×512 の画像を入力として受け取って 768 次元の特徴量を出力する．wd-vit-tagger-v3 と同様にヘッドを除き，最終層から 5 番目の Transformer ブロックの出力を後続の MLP への入力とした．したがってパッチ数 $32 \times 32 = 1024$，出力次元 768 から，$(1024, 768)$ が MLP への入力サイズとなる．

MLP

CVC-Tagger と CVC-SigLIP では同じ構成の MLP を用いた．MLP は画像エンコーダーから出力された画像パッチシーケンスを入力として受け取り，シーケンスを維持したままパッチごとに特徴量を変換する．MLP は全結合層と活性化関数から成り，画像エンコーダーの出力する特徴量 768 次元を隠れ次元 2048 の特徴量に変換したのち活性化関数 Swish [30] を適用し，再び全結合層を用いて 768 次元に変換する．

© 1959 Information Processing Society of Japan

3.3 学習設定

表 1 に本実験における学習設定を示す．CVC-Tagger と CVC-SigLIP では同じ学習設定を用いた．最適化関数は RAdamScheduleFree [31] を用い，学習率は 1e-4，バッチサイズは 16 で 20000 ステップ学習した．

Irodori-TTS [8] では，推論時に Classifier-Free Guidance (CFG) [32] を使用する．CFG では，条件付き予測と unconditional 条件による予測を組み合わせるため，unconditional 条件の定義が出力に影響する．Irodori-TTS の実装では，推論時の unconditional な参照条件は Attention Mask を全て 0 にし，参照特徴への Attention を遮断することで実現されている．この実装を CVC にそのまま適用すると，画像特徴を参照する条件付き予測と，画像条件用の Attention を完全に無効化した予測を組み合わせることになる．一方で，IP-Adapter [7] では，推論時の CFG において，unconditional 条件として空の画像の代わりにノイズ画像を用いると出力品質が向上することが報告されている．[*3] そこで本研究では，画像条件用の Attention は有効にしたまま，参照画像特徴をノイズ画像から抽出した特徴量に置き換え，これを unconditional 条件として用いた．この設定では，学習時に特定の空条件を学習させる必要がないため，CFG のための画像条件ドロップアウトは行わなかった．

学習時のロス関数 $\mathcal{L}_{\text{img}}$ は音声参照時の損失関数の式 (2) を基に，参照音声条件 y_{ref} を画像条件 y_{img} に置き換えたものであり，以下のように表される．

$$\mathcal{L}_{\text{img}} = \mathbb{E}_{t,x_0,x_1} \| v_t - v_\theta(x_t, t, y_{\text{text}}, y_{\text{img}}) \|_2^2 \qquad (3)$$

CVC-Tagger と CVC-SigLIP では，同じ学習設定を用いて $\mathcal{L}_{\text{img}}$ を最小化するように学習を行った．学習は 1 台の NVIDIA GeForce RTX 4070 Ti SUPER を用いて，それぞれ約 5 時間かけて行われた．

表 1: 提案手法の学習設定

	CVC-Tagger	CVC-SigLIP
アーキテクチャ		
エンコーダー	wd-vit-tagger-v3[25]	SigLIP2-B/16[28]
出力層	最終層から 5 層目	
入力画像サイズ	448×448	512×512
MLP 隠れ次元	2048	
MLP 層数	1	
MLP 活性化関数	Swish [30]	
ドロップアウト	0	
学習設定		
最適化関数	RAdamScheduleFree[31]	
Adam β_1, β_2	0.9, 0.99	
Weight decay	0.01	
学習率 (LR)	1e-4	
LR scheduler	なし	
バッチサイズ	16	
学習ステップ数	20000	
最大テキスト長	32	

4. 評価

本実験では定性評価と定量評価を実施した．定性評価では 6 人の評価者を用いて，画像と生成された音声の対応性を評価した．定量評価では UTMOSv2 [33] を用いた音声自然性の評価と，複数のマルチモーダル LLM を用いた画像と音声の対応性の評価を実施した．

本評価では，同一のキャラクター画像および同一の発話テキストに対して，声質条件の与え方が異なる複数の音声合成条件を比較した．比較対象は，参照する声質条件を与えない Irodori-TTS[8] (No reference)，テキストキャプションにより声質を指定する Voice Design[20]，そして提案手法 CVC-Tagger と CVC-SigLIP の 4 条件である．

4.1 評価用データセット

本評価では提案手法の有効性を検証するため，学習データセットに含まれない，評価用に新規生成した架空キャラクター画像を用いた．評価対象のキャラクターには対応する実在話者が存在しないため，Ground truth となる音声は存在しない．

キャラクター画像は NovelAI Diffusion V4.5 Curated [34]，AnimagineXL 4.0 [35]，ChatGPT-Images-2.0 [36] を用いて手動で生成した．内訳は女性キャラクターが 11 件，男性キャラクターが 12 件，動物や怪物，ロボットなどの非人間キャラクターが 7 件の合計 30 キャラクターで，1 キャラクターにつき 1 枚の画像を用意した．このとき，イメージされるキャラクター性が多様となるように，髪型や服装，表情，ポーズ，テーマなどを変えながら生成した．

評価単位は，1 枚のキャラクター画像と 1 つの読み上げテキストからなる画像・セリフペアである．各キャラクターに対して，キャラクター固有セリフ 10 件と，全キャラクターで共通するナレーション用セリフ 7 件を用意した．

キャラクター固有セリフは LLM を用いて合成した．Gemini 3.1 Pro Preview [37] を用いて，キャラクター画像を条件に，そのキャラクターが話しそうなセリフ・その時の声の特徴を説明するテキストキャプション (Voice Caption)・そのセリフを話している時の状況説明を 10 組ずつ生成した．このときの Voice Caption はセリフごとに異なり，Voice Design 条件の入力として使用された．また，キャラクター固有の口調や雰囲気を含まない，物語のナレーション形式となる共通セリフ 7 件を用意した．加えて，それぞれのキャラクターごとに汎用的な Voice Caption を 1 つずつ生成し，ナレーション用セリフを読み上げる際の Voice

[*3] https://github.com/tencent-ailab/IP-Adapter/issues/51

© 1959 Information Processing Society of Japan

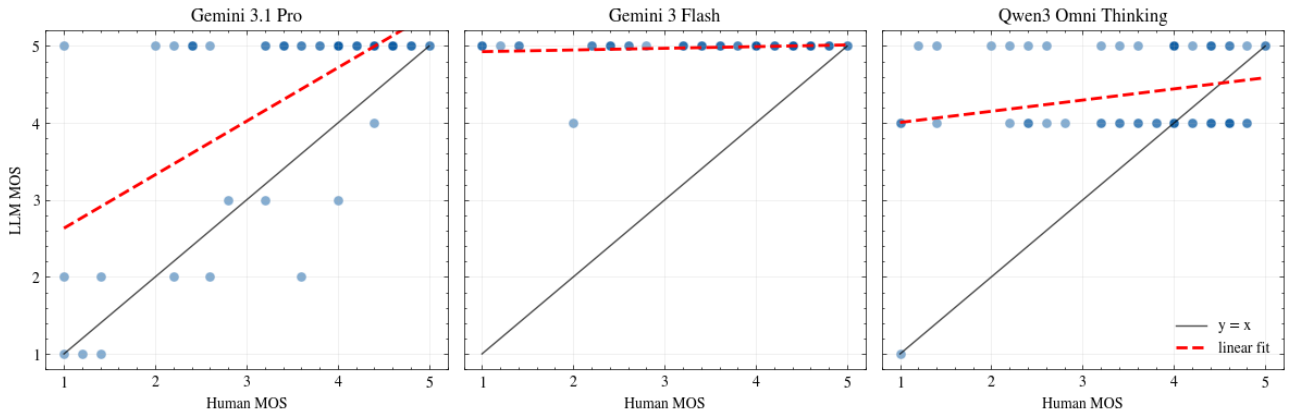


図 3: 人間による親和性 MOS とマルチモーダル LLM による親和性 MOS のスコア散布図

表 2: 有人評価における親和性 MOS 評価 (95% 信頼区間)

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

表 3: 自然性 MOS 評価 (95% 信頼区間，高いほど良い)

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

Design 条件の入力として使用した．

以上により，各キャラクターは 10 件の固有セリフと 7 件の共通セリフを持ち，全体では $30 \times (10+7) = 510$ 件の画像・セリフペアが構成される．これらのペアについて，No reference，Voice Design，CVC-Tagger，CVC-SigLIP の 4 条件それぞれ 510 件の音声を生成し，合計 2040 件の音声を評価用として用意した．

4.2 親和性の有人評価

定性評価では，作成した評価データセットのサブセットを用いて，音声が画像とどれくらい親和性があるかを評価する有人評価アンケートを実施した．本アンケートでは 6 人の評価者を用いて，画像と生成された音声の親和性を 5 段階 Mean Opinion Score (MOS) (1: 全く合っていない，5: 非常によく合っている) で評価した．有人評価では評価者の負担を考慮して，すべての音声を評価する代わりに評価データセットの一部をサブセットとして用いて評価した．サブセットは，30 キャラクターのうち手動で選定された 14 キャラクターについて，キャラクター固有セリフのうちランダムに選定された 1 セリフずつのみを用いて構成された．内訳は女性キャラクター 4 件，男性キャラクター 6 件，非人間キャラクター 4 件である．評価者には，これらの 14 キャラクターと 4 条件 からなる 56 件の画像・音声・セリフの三つ組ペアについてランダムな順序で 1 ペアずつ提示し，画像と生成された音声の親和性を 5 段階評価で評価してもらった．

表 2 に人間の評価者による 95% 信頼区間付き親和性 MOS スコアを示す．人間による評価では，提案手法 CVC-SigLIP が最も高い MOS スコア 4.400 ± 0.200 を獲得し，次いで CVC-Tagger が 4.029 ± 0.279 となり，既存手法である Voice Design 3.186 ± 0.636 と比較して提案手法が上回った．

興味深いことに，一切声質の参照がない No reference 条件でさえも 2.514 ± 0.507 というスコアが得られた．実際に，役割語や口調によってキャラクターの役割がセリフ単体で表現されている場合，何も追加の条件を与えなくても，キャラクター画像とある程度対応した声質で生成されるケースが存在した．これは大規模に事前学習された Irodori-TTS [8] が，既に多様なキャラクター性のある声質表現を獲得しており，Voice Caption や CVC はそれを適切に引き出す役割を果たしていることを示唆している．一方で，キャラクター共通で用意したナレーションのセリフは個別のキャラクターイメージが反映されないため，そのような現象は見られなかった．ただし，有人評価用サブセットにはこれらのナレーションセリフが含まれなかったため，このようなセリフを有人評価の対象に含めると，No reference 条件の親和性スコアはさらに下がる可能性がある．

4.3 UTMOSv2 による音声の自然性評価

UTMOSv2 [33] を用いて，評価データセットで生成した音声の自然性を定量的に評価した．UTMOSv2 は，音声の自然性を客観的に評価するためのシステムであり，音声の自然性を 1 から 5 のスケールで評価する．

それぞれの条件ごとに 510 件の音声を評価した平均 UTMOSv2 スコアと 95% 信頼区間を **表 3** に示す．

© 1959 Information Processing Society of Japan

表 4: 人間の評価者と MLLM 評価者の相関係数

MLLM	Pearson r	Spearman ρ	Kendall τ
Gemini 3.1 Pro	0.628	0.514	0.434
Gemini 3 Flash	0.180	0.172	0.145
Qwen3 Omni Thinking	0.267	0.155	0.131

提案手法である CVC-Tagger が最も高い自然性スコア 2.651 ± 0.039 を獲得した．また，CVC-SigLIP の自然性スコアは 2.556 ± 0.037 であり，No reference 条件の 2.593 ± 0.043 と比較しても自然性を大きく損なっていないことが示唆される．

4.4 マルチモーダル LLM による親和性評価

近年，有人評価の代わりに LLM を評価者として用いる LLM-as-a-Judge 手法が注目されている．そこで本研究では，画像・音声・テキストを扱えるマルチモーダル LLM (MLLM) を用いた評価を検討した．

MLLM による評価では有人評価と同様に 5 段階で画像と音声の親和性評価をした．評価者として用いた MLLM は Gemini 3.1 Pro [37]，Gemini 3 Flash [38]，Qwen3-Omni-30B-A3B-Thinking [39] の 3 つである．

図 3 は，MLLM が評価したうち，有人評価に用いたサブセット 56 件における，有人評価の親和性 MOS と MLLM による親和性 MOS の散布図であり，x 軸が有人評価の親和性 MOS，y 軸が MLLM による親和性 MOS を表す．図 3 から、Gemini 3 Flash と Qwen3 Omni Thinking は評価が甘く，人間が低いスコアを与えたサンプルに対しても高いスコアを与える傾向が窺える．

有人評価の親和性 MOS と MLLM の親和性 MOS の相関を調べるため，Pearson の相関係数 r[40]，Spearman の順位相関係数 ρ[41]，Kendall の順位相関係数 τ[42] をそれぞれ計算した．**表 4** に結果を示す．Pearson の r はスコア値そのものの線形な対応を，Spearman の ρ および Kendall の τ は順位レベルでの対応を表す．Gemini 3.1 Pro は Pearson 相関 $r = 0.628$，順位相関 $\rho = 0.514$，$\tau = 0.434$ と中程度の正の相関を示した．他の MLLM 評価者の人間評価との相関はさらに低い．したがって，本実験設定では，MLLM による MOS 評価は人間評価の傾向を一貫して再現しているとはいえない．これらの結果から，MLLM を人間の評価者の完全な代替として利用することには課題があると言える．

5. おわりに

本研究では，キャラクター画像を入力として声質スタイルを制御する日本語テキスト音声合成手法 Character Voice Control を提案した．画像・音声・発話テキストを対応付けた三つ組データセットを構築し，事前学習済み日本語 TTS モデルである Irodori-TTS [8] に画像エンコーダーと MLP を統合することで，キャラクター画像を条件とした音声合成を実現した．

評価実験では，提案手法である CVC-SigLIP が有人評価において最も高い親和性 MOS 4.400 ± 0.200 を獲得し，CVC-Tagger も Voice Design を上回る評価を得た．また，UTMOSv2 による自然性評価では，CVC-Tagger が最も高いスコアを示し，提案手法は音声の自然性を大きく損なうことなく，画像に基づく声質制御を実現できることが確認された．さらに，MLLM を用いた評価を検討した結果，Gemini 3.1 Pro は人間の評価と一定程度近い傾向を示した．一方で，MLLM は人間が低く評価したサンプルにも高い評価を与えやすく，細かな声質の違いを十分に捉えられない場合があることが明らかになった．

今後の課題として，まず学習データセットの規模と多様性の拡充が挙げられる．本研究で用いたデータセットは短い発話が中心であり，長文生成時にベースラインよりも途中で音声が途切れる例が見られた．長い発話を含むデータで学習することで，長文入力に対する発話完了率を改善できる可能性がある．加えて，本研究では MLP のみを学習対象としたが，RFDiT など他モジュールの凍結解除や画像特徴量の統合方法を改善することで，より精密な声質制御が期待できる．また，本研究では Irodori-TTS をベースとして提案手法を実装したが，他の TTS モデルや異なる音声生成アーキテクチャに対しても同様の画像条件付き声質制御が適用可能であるかを検証する必要があるだろう．さらに，画像と声質の対応関係には主観性や文化的なバイアスが含まれるため，人間の感性に沿った定量的な評価手法の確立も重要である．特に，MLLM による評価は依然として人間の評価との乖離が見られたため，有人評価を補完しつつ画像と声質の対応性を安定して測定できる評価指標が求められる．

謝辞 本実験の評価の一部にて，電気通信大学人工知能先端研究センター (AIX) の計算機を利用した．

参考文献

[1] Seedance, T., Chen, D., Chen, L. et al.: Seedance 2.0: Advancing Video Generation for World Complexity (2026).
[2] Hu, H., Zhu, X., He, T. et al.: Qwen3-TTS Technical Report, *arXiv preprint arXiv:2601.15621* (2026).
[3] Zhu, H., Ye, L., Kang, W., Yao, Z., Guo, L., Kuang, F., Han, Z., Zhuang, W., Lin, L. and Povey, D.: OmniVoice: Towards Omnilingual Zero-Shot Text-to-Speech with Diffusion Language Models, *arXiv preprint arXiv:2604.00688* (2026).
[4] Zhou, Y., Zeng, G., Liu, X., Li, X., Yu, R., Wang, Z., Ye, R., Sun, W., Gui, J., Li, K., Wu, Z. and Liu, Z.: VoxCPM: Tokenizer-Free TTS for Context-Aware Speech Generation and True-to-Life Voice Cloning,

© 1959 Information Processing Society of Japan

arXiv preprint arXiv:2509.24650 (2025).
[5] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J. and Rombach, R.: SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis (2023).
[6] Zhang, L., Rao, A. and Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models (2023).
[7] Ye, H., Zhang, J., Liu, S., Han, X. and Yang, W.: IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models (2023).
[8] Arata, C.: Irodori-TTS: A Flow Matching-based Text-to-Speech Model with Emoji-driven Style Control, `https://github.com/Aratako/Irodori-TTS` (2026).
[9] Goto, S., Onishi, K., Saito, Y., Tachibana, K. and Mori, K.: Face2Speech: Towards Multi-Speaker Text-to-Speech Synthesis Using an Embedding Vector Predicted from a Face Image, *Proc. Interspeech 2020*, No. 1321-1325 (2020).
[10] Wan, L., Wang, Q., Papir, A. and Moreno, I. L.: Generalized End-to-End Loss for Speaker Verification (2020).
[11] 菊池遥斗, 能勢 隆, 小林清流, 林崎 由, 橋本 佳, 伊藤彰則: F2S-SBV2：任意のアニメ調キャラクター顔画像に適した話者性を有するテキスト音声合成の検討, 情報処理学会研究報告, pp. 1–7 (2025).
[12] 菊池遥斗, 能勢 隆, 伊藤彰則: キャラクター画像からの音声合成に向けた動向分析, 情報処理学会研究報告, pp. 1–7 (2024).
[13] 菊池遥斗, 能勢 隆, 小林清流, 林崎 由, 橋本 佳, 伊藤彰則: アニメ調キャラクター顔画像に適するクロスモーダル音声合成のための再現性・親和性改善の検討, 日本音響学会 2026 年春季研究発表会講演論文集, pp. 1109–1112 (2026).
[14] Darefsky, J.: Echo-TTS (2025).
[15] Liu, X., Gong, C. and Liu, Q.: Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow (2022).
[16] Peebles, W. and Xie, S.: Scalable Diffusion Models with Transformers (2023).
[17] Arata, C.: Irodori-TTS: A Flow Matching-based Text-to-Speech Model with Emoji-driven Style Control, `https://huggingface.co/Aratako/Irodori-TTS-500M-v2` (2026).
[18] Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.-Y., Chuang, C.-Y. et al.: Movie gen: A cast of media foundation models, *arXiv preprint arXiv:2410.13720* (2024).
[19] Arata, C.: Semantic-DACVAE-Japanese-32dim: Lightweight Audio VAE for Japanese Speech, `https://huggingface.co/Aratako/Semantic-DACVAE-Japanese-32dim` (2026).
[20] Arata, C.: Irodori-TTS: A Flow Matching-based Text-to-Speech Model with Emoji-driven Style Control, `https://huggingface.co/Aratako/Irodori-TTS-500M-v2-VoiceDesign` (2026).
[21] Team, S.: Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier, `https://github.com/snakers4/silero-vad` (2024).
[22] SYSTRAN: Faster Whisper transcription with CTranslate2, `https://github.com/SYSTRAN/faster-whisper` (2023).
[23] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I.: Robust Speech Recognition via Large-Scale Weak Supervision (2022).
[24] McFee, B., McVicar, M., Faronbi, D. et al.: librosa/librosa: 0.11.0 (2025).
[25] SmilingWolf: WD ViT Tagger v3, `https://huggingface.co/SmilingWolf/wd-vit-tagger-v3` (2024).
[26] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2021).
[27] Ryu, S., Pengqi, L., Martín Juan, J. and de Prado Alonso, I.: F Lite Technical Report (2025).
[28] Tschannen, M., Gritsenko, A., Wang, X., Naeem, M. F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmsen, J., Steiner, A. and Zhai, X.: SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features (2025).
[29] timm: Model card for vit_base_patch16_siglip_512.v2_webli, `https://huggingface.co/timm/vit_base_patch16_siglip_512.v2_webli` (2025).
[30] Ramachandran, P., Zoph, B. and Le, Q. V.: Searching for Activation Functions (2017).
[31] Defazio, A., Yang, X. A., Mehta, H., Mishchenko, K., Khaled, A. and Cutkosky, A.: The Road Less Scheduled (2024).
[32] Ho, J. and Salimans, T.: Classifier-Free Diffusion Guidance (2022).
[33] Baba, K., Nakata, W., Saito, Y. and Saruwatari, H.: The T05 System for The VoiceMOS Challenge 2024: Transfer Learning from Deep Image Classifier to Naturalness MOS Prediction of High-Quality Synthetic Speech, *IEEE Spoken Language Technology Workshop (SLT)*, pp. 818–824 (online), DOI: 10.1109/SLT61566.2024.10832315 (2024).
[34] Anlatan: NovelAI Diffusion V4.5 Curated is Here!, `https://blog.novelai.net/novelai-diffusion-v4-5-curated-is-here-9e5ba1f628f3` (2025).
[35] Cagliostro Research Lab: Animagine XL 4.0, `https://huggingface.co/cagliostrolab/animagine-xl-4.0` (2025).
[36] OpenAI: Introducing ChatGPT Images 2.0, `https://openai.com/index/introducing-chatgpt-images-2-0/` (2026).
[37] Google: Gemini 3.1 Pro Preview, `https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview` (2026).
[38] Google: Gemini 3 Flash, `https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview` (2026).
[39] Xu, J., Guo, Z., Hu, H. et al.: Qwen3-Omni Technical Report (2025).
[40] Pearson, K.: VII. Note on Regression and Inheritance in the Case of Two Parents, *Proceedings of the Royal Society of London*, Vol. 58, No. 347–352, pp. 240–242 (online), DOI: 10.1098/rspl.1895.0041 (1895).
[41] Spearman, C.: The Proof and Measurement of Association Between Two Things, *The American Journal of Psychology*, Vol. 15, No. 1, pp. 72–101 (online), DOI: 10.2307/1412159 (1904).
[42] Kendall, M. G.: A New Measure of Rank Correlation, *Biometrika*, Vol. 30, No. 1/2, pp. 81–93 (online), DOI: 10.1093/biomet/30.1-2.81 (1938).

© 1959 Information Processing Society of Japan