声質選択インタフェースを用いた音声合成シナリオ執筆支援システム

滝田 巧平[†1]　青柳 西蔵[†1]　平井 辰典[†1]

概要：書き上げたシナリオを読み上げることでセリフの自然性や個性のブレを確認することができる．この読み上げを自身で行うのではなく，登場キャラクタのイメージに近い声で行うことができれば，よりセリフのイメージがつきやすくなると考えられる．しかし，従来の音声合成システムでは番号や話者名等によって声を選ぶ必要があり，一目で声を判別することが難しい．そこで，本研究では学習に用いた話者の声を二次元平面上に可視化することで，イメージ通りの声を選択しやすくするインタフェースを提案する．このインタフェースを搭載したシナリオ執筆支援システムを用いてシナリオを執筆してもらうことで本システムによる執筆者への影響を評価する．その結果，イメージ通りの声を選択することには一部成功し，キャラクタのイメージが想像しやすくなることで，執筆活動に貢献できることを示す．

1. はじめに

シナリオを執筆する際には，シナリオが書き上がるとシナリオチームのメンバーが集まって書いたシナリオを読み上げ演技をしてみる作業が制作の過程に含まれる場合もある [1]．ライター以外の人でも，自身が書いた文章を実際に読み上げて違和感や間違いがないか確認した経験があるのではないだろうか．一方シナリオに関しては一般的な作文とは異なり，キャラクタが数多く登場し，各キャラクタに性別や性格や話し方などの設定がされている．それに加えて，言語化が難しいキャラクタの声に関してもライターの中ではイメージされていると考えられる．もし，ライターのイメージ通りの声で読み上げを行うことができれば，ライター自身が読み上げるよりもセリフの自然性やキャラクタの個性のブレがないかを確認することができてより良いシナリオ執筆に役立つのではないかと考えられる．

しかし，イメージ通りの声を探すことは難しい．例えば，VOICEROID[2] のようなテキスト読み上げシステムでは，話者を選択する際に話者に振られている ID（名前）を目印に選択する仕様となっており，ID のみを頼りに目当ての声を探さなければならない．また，Jia ら [3] の参照音声から抽出した話者特徴量を音声合成モデルで用いることでターゲットに近い声を合成する手法があるが，このような手法の場合にはターゲットの音源ファイルを用意する必要がある．

そこで本研究では，二次元平面上で声の位置関係が視覚化され，ライターがキャラクタのイメージに近い音声を直感的に探せるインタフェースを開発する．さらに，そのインタフェースを搭載したシナリオ執筆支援システムを用いて実際にシナリオを執筆してもらうことでライターの執筆活動への影響を評価する．

2. 関連研究

シナリオの執筆やコンテンツの制作を支援するアプローチとしてテキスト情報でのアプローチと音声情報でのアプローチがある．以下に，各アプローチについて記述する．

まず，シナリオの執筆支援のアプローチとしてテキストの自動生成が挙げられる．近年では OpenAI 社が開発した ChatGPT を始めとする大規模言語モデルによって人が書いたものと遜色ないテキストを生成でき，これらの技術はライターの執筆の支援に役立つと考えられる．

次に，シナリオの執筆支援のアプローチとして文章校正技術が挙げられる．鈴木らの構築した執筆支援システム [4] では，深層学習モデルの Encoder-Decoder モデルによる文章生成アプローチによって執筆者が記述した文の文脈を考慮して修辞表現を付加した候補文を提案することで文章表現の推敲を支援する．文章校正技術は Microsoft 社が提供する Microsoft Word のような製品にも用いられていてライターの執筆に役立っていると考えられる．

さらに，シナリオ執筆を行う上で，書き進めるごとに情報が増えていくシナリオの情報を管理することが重要である．戀津ら [5] が提案したシステムでは，シナリオを執筆する際に発生する情報を，PHP と MySQL のデータベース

[†1] 駒澤大学

によるシステムによって管理，表示を行い，シナリオライターの情報整理を補助することでシナリオ執筆を支援する．

以上に挙げたシナリオ執筆支援技術は，テキストの自動生成・自動校正・情報管理などによってライターの支援を行うものである．一方本研究で提案するシステムは，ライターが書き上げたシナリオをイメージに近い声の合成音声で読み上げることによってシナリオやキャラクタのイメージの具体化を支援することを目的とし，シナリオにおけるキャラクタやキャラクタのセリフの作成に注目する．

本研究と同様に音声情報を用いることによってシナリオ執筆を支援するためのアプローチもいくつか考えられる．

まず，シナリオ執筆に役立つシステムとして VOICEROID[2] のような音声読み上げソフトがある．これらの読み上げソフトは，テキストの読み上げ，アクセントや声の高さなどの調整，音声ファイルなどの書き出しなどを行うことができる．音声読み上げソフトによって書き上げたテキストを読み上げることで文章の間違いなどを確認することに使える．合成した音声の Wav ファイル書き出しも行えるため合成した音声自体を朗読や実況などのコンテンツとして使用することもできる．

音声から画像を生成するマルチモーダルな研究キャラクタ制作にを利用した研究もある．栗山ら [6] は Speech2Face という手法を用いて音声からキャラクタの顔を生成することによって対話システム開発時のキャラクタの決定補助に役立てている．

音声とキャラクタの関係についての研究事例として，声優推薦ツールに関する研究もある．酒井ら [7] はポータブルゲームの音声を収録した音声データベースを構築し，音声データから算出した音響特徴量と被験者の回答から得た音声データの印象値を学習させることで未知のキャラクタに対する印象値から適切な音響特徴量を推定し，推定された音響特徴量を元に適した声優候補リストを生成する．さらにウェブ上の文書からキャラクタ間の距離を算出して可視化する機能も提案しており，これら二つの機能によって声優の推薦を目指している．声優の推薦を行うことでゲームキャラクタの声優決めの支援などが可能となる．

このように音声を基に自動的にキャラクタを生成する研究や最適な声優を推薦する研究などがあるが，本研究は，音声合成による読み上げによってライターのシナリオ執筆やキャラクタのイメージの具体化を支援し，ライターの創作活動を支援する．音声合成を用いたテキスト読み上げシステムは様々にあるが．本研究では，ライター自身がイメージ通りの声を二次元インタフェースから選択できるような機能の実現を目指す．

3. シナリオ執筆支援システム

本研究では，シナリオに登場するキャラクタのイメージに近い声をライターが選択できる二次元平面インタフェースを搭載したシナリオ執筆支援システムを開発する．以下に開発したシステムの詳細について記述する．

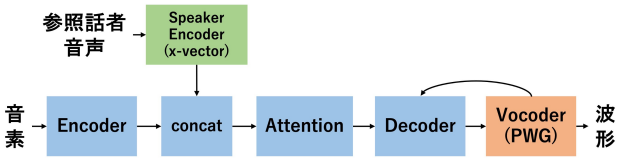

図 1 音声合成モデルの構成
Fig. 1 architecture of tts-model

3.1 音声合成モデル

近年の音声合成技術で合成される音声は自然音声と見分けるのが難しいほど高品質な合成を実現している．これを実現しているのは深層学習に基づく音声合成モデルを使用しているからである．ニューラルネットを用いた一貫学習に基づく音声合成モデルの流れでは，テキストからメルスペクトログラムなどの音響特徴量を生成する音響モデルと音響特徴量から音声波形を生成するボコーダの二つのモデルで構成される．ここでは，提案システムで使用した二つのモデルについて詳述する．**図 1** に本システムに用いる音声合成モデルの構成図を示す．

まず，音素からメルスペクトログラムを生成するための音響モデルには，Tacotron2[8] に用いられている音響モデルを使用した．Tacotron2 の音響モデルはメルスペクトログラムの残差を予測する Post-Net と呼ばれるモジュールを含んだ注意機構 (Attention) 付き sequence-to-sequence モデルである．この音響モデルを用いることで音素からメルスペクトログラムへの変換を行う．この時，話者認証モデルである x-vector[9] を用いて音声から抽出した話者特徴量を音響モデルの Encoder の出力に結合してから Attention 及び Decoder に入力することで，参照話者に近い音響特徴量を生成することができる．Encoder の出力に話者特徴量を結合するモデルの構造は，Jia らの研究 [3] を参考にし，話者特徴量を Encoder の出力に結合する前に 512 次元の話者特徴量を全結合層によって 64 次元する構造は Cooper らの研究 [10] を参考に実装した．

メルスペクトログラムから音声波形を生成するニューラルボコーダには Parallel WaveGAN[11] を用いた．Parallel WaveGAN は非自己回帰型の WaveNet に敵対的生成ネットワークを用いることで高速に品質の高い音声を合成できるモデルである．敵対的学習プロセスの安定性と効率を向上させるために補助的な損失関数として多重解像度 STFT loss が用いられている．

Tacotron2 及び Parallel WaveGAN の学習には JSUT コーパス [12] 及び JVS コーパス [13] を用いた．JSUT コーパスからは 4700 発話分 JVS コーパスからは 118 発話分のデータを 100 話者分用い，計 11800 発話を学習に利用

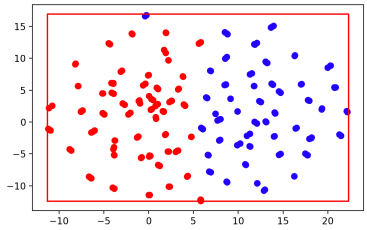

図 2 UMAP によって可視化された全話者（丸）とその範囲（線）
Fig. 2 The Speakers and The Range Visualized by UMAP

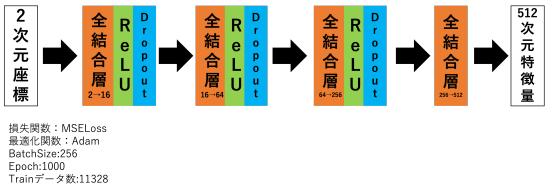

図 3 2 次元から 512 次元の変換
Fig. 3 Conversion from 2D to 512D

した．また，Tacotron2 で使用した話者特徴量を抽出する x-vector は Hamada[14] らが提供している学習済みモデルを用いた．

3.2 話者選択用二次元インタフェース

本節では，イメージ通りの声を選択するための二次元平面インタフェースについて記述する．

本研究で提案するインタフェースは，ライターがキャラクタのイメージに近い音声を選択しやすくするために，音声合成モデルの学習に用いた各話者の声を二次元平面上にプロットしている．例えば，**図 2** に示す 2 次元平面は，シナリオ執筆支援システムに搭載している話者選択のための二次元平面インタフェースを抜き出したものである．女性話者は赤い丸でプロットされており，男性話者は青い丸でプロットされており，それぞれ二次元平面上の左右に分かれている．そして，この二次元平面上の一点を押下することで合成する声を選択することができる．このプロットは次元削減手法の UMAP[15] を用いて各話者の音声データから抽出した話者特徴量を基に 2 次元にプロットしている．この UMAP によって似ている話者が近く，似ていない話者が遠くへ配置されており，ライターは直感的に声を探すことができる．

前述した音声合成モデルでは，話者特徴量を参照話者音声から取得していたが，このインタフェースは，押下した位置の 2 次元座標を取得し，その座標から話者特徴量を取得する．二次元座標から話者特徴量を取得するために，以下の二つの手法のどちらか一方をシナリオ執筆支援システムの選択欄によって切り替えて用いる．

一つ目の手法はニューラルネットワークを用いる手法である．まず，x-vector によって JVS コーパスの 11328 個の音声データから 512 次元の話者特徴量を抽出する．この話者特徴量を教師データとする．512 次元の話者特徴量データを UMAP[15] を用いて 2 次元座標のデータに削減する．そして，この 2 次元のデータを入力データとして，2 次元から 512 次元への変換を行うニューラルネットワークを学習し，2 次元座標から話者特徴量を復元できるようにする．このニューラルネットワークのネットワーク図は**図 3** に示す．

二つ目の手法はニューラルネットワークを用いない手法である．2 次元のインタフェース上で押下された二次元座標 (x, y) と UMAP で得た学習済み話者の座標データセット (X, Y) の中で最も近い座標をユークリッド距離を基に算出する．以下の式で得られた最小の d に対応する座標がインタフェース上で押下された二次元座標 (x, y) に最も近い座標である．

$$d = \sqrt{(X - x)^2 + (Y - y)^2} \tag{1}$$

二つ目の手法では，最も近い座標に対応する話者特徴量を話者特徴量データセットから取り出して音響モデルの条件付けとする．

3.3 執筆支援システムの諸機能

前述した音声合成モデル及びインタフェースを用いたシナリオ執筆支援システムの全体像と各機能について記述する．

本システムの全体像を**図 4** に示す．本システムの左側は音声合成に関連した機能，右側にはシナリオ執筆に関連した機能が配置されている．

初めに，左側の音声合成に関連した機能について記述する．左上のテキスト入力欄は選択した話者の声を確かめるために使用する．任意のテキストを入力欄に記述し，下に配置された合成ボタンを押下することで入力欄に書かれたテキストを読み上げる音声が合成される．合成ボタンの下に位置する 2 次元平面は前述した話者選択インタフェースである．本インタフェース上の任意の点を押下することで 3.2 節で述べた手法によって話者特徴量が取得される．話者特徴量が取得された後，利用者への確認として定められた 5 文字のテキストを読み上げた音声が取得された特徴量を用いて合成される．この二次元平面は，x 座標が-11 から 22，y 座標が-12 から 16 を前後の範囲で設定されている．これは，学習に用いたデータの 2 次元座標の最大・最小の値を基に設定している．

次に，右側のシナリオ執筆に関連した機能について記述する．緑色で示されている箇所がシナリオが表示される箇所で，その下に配置されたボタン及び入力欄によって各キャラクタのセリフを入力できる．人物登録ボタンを押下するとシナリオに登場するキャラクタの登録ができる．人物登録ボタンによってキャラクタを登録するとボタンの下

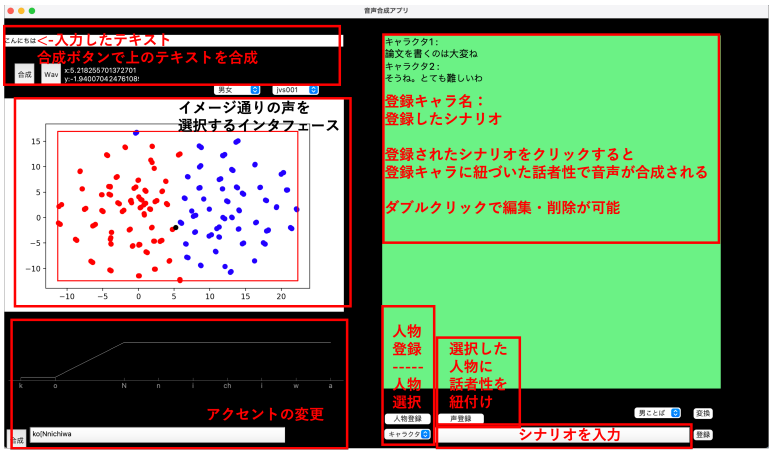

図 4 提案システムの全体像
Fig. 4 overall image of the proposed system

にある選択欄から登録されたキャラクタを選択することができる．登録したキャラクタを人物選択欄から選び，左側の話者選択用インタフェースで声を選んだ状態で声登録ボタンを押下すると選択中のキャラクタに声を紐付けることができる．人物選択欄でキャラクタを選び，その隣に位置するテキスト入力欄にセリフを入力してから登録ボタンを押下することで緑色のシナリオ欄にシナリオが追加されていく．そして，シナリオ欄の各シナリオを押下すると登録したキャラクタに紐付いた声でセリフが読み上げられる．また，ダブルクリックをすることで各セリフの編集と削除を選択するウィンドウが表示され，ライターは登録済みのシナリオの編集や削除を行うことができる．

4. 評価実験

開発したシナリオ執筆支援システムを用いて被験者にシナリオを執筆してもらうことで執筆活動にどのような影響をもたらすのか評価する．

本インタフェースに関する二つのことについての評価を目的として評価実験を行う．一つ目の目的は，本インタフェースによってライターがイメージに近い声を選択できるかの評価である．二つ目は，本システムを用いてイメージに近い声でシナリオを読み上げることで，キャラクタの個性のブレに気づいたり，セリフが自然でないと気づいたりといった，執筆作業に良い影響をもたらすかの評価である．前者を音声の評価とし，後者を執筆実験と呼称する．

4.1 二次元座標から復元して合成した音声の評価

執筆実験を行う前に、本研究で構築したニューラルネットワークで復元された話者特徴量を用いることで学習に用いた話者の元の音声を合成できるかを評価する．最初に，9 人の被験者に対して，3 種類の音声のみが流れる動画を Google フォームを用いて提示する．提示された動画は次の 3 種である．

表 1 自然音声・素の特徴量・座標で合成した音声類似度の比較

	自然音声	通常の特徴量で合成	座標から復元・合成
男女	4.57	3.41	2.46
男性	4.33	3.59	2.26
女性	4.81	3.22	2.67

- 同じ自然音声が 2 回流れる動画
- 教師データの話者特徴量を用いた合成音声と教師データに対応する元々の話者の自然音声が流れる動画
- 二次元座標から復元した話者特徴量を用いた合成音声と座標に対応する元々の話者の自然音声が流れる動画

この 3 種類の動画を各被験者に 6 話者（男性 3 話者・女性 3 話者）分提示した．被験者は提示された音声を聞いて，類似性に焦点を当てて評価を行う．被験者は提示された各動画から流れる二つの音声の類似性を 1 から 5 までの 5 段階で評価する．評価値は，1 が非常に似ていないことを示し，5 が非常に似ていることを示す．**表 1** には評価の結果を男女混合・男性話者・女性話者の三つに分けて平均値で示す．

評価の結果，座標から合成した音声は自然音声や通常の話者特徴量から合成した音声と比べて低い評価値となった．この理由として，2 次元の座標から 512 次元の話者特徴量を復元の過程で通常の話者特徴量に含まれている情報が損なわれてしまったことで，全体的に同じような音声が合成されてしまい，元の自然音声の特徴を捉えきれなかったからだと考えられる．

4.2 執筆実験の内容

執筆実験ではシナリオや小説や漫画などの物語を執筆した経験のある被験者 2 人とそれらの執筆経験がない被験者 2 人の計 4 人に協力してもらった．4 人の被験者には開発したシステムの使い方を 10 分ほど説明した後，実際にシステムを操作して慣れてもらった．被験者がある程度システムに慣れたと感じたら，被験者にシステムを用いて短い

表 2 シナリオ執筆実験後の聞き取り調査結果

名前 (執筆経験)	1 イメージ通りの声の生成	2 セリフの修正	3 書きやすさ	4 普段との違い
被験者 K(有)	ナレーションは納得 桃太郎のイメージと 老人の声が合成できず	特になし ボイスレパートリー が多ければ	ボイスを聞けることで 感じたが合成に スピード感が足りなかった	そこまでなかった テンポ感が失われた
被験者 W(有)	若い女の子の声は納得 老人の声合成できず	セリフの修正はなかったが 自分のイメージ像と 結びつきやすくなり 書きやすくなった	書きやすさを感じた	自分と 相手のイメージの ズレを少なくできそう
被験者 H(無)	女性の声が特に納得	セリフの修正なし	キャラクタの声を 決めたことで キャラクタ性の維持できた	キャラクタ独自の話し方や 方言を意識して書いてた
被験者 A(無)	イメージに近いもので納得	違和感なし	声をつけることで キャラクタの設定 がつけやすく書きやすかった	イメージが想像しやすく 次の展開を文章で 考えやすかった

シナリオを書いてもらった．この時，被験者が書きたいシナリオの題材がない場合，スムーズに執筆ができるように執筆を始める前に簡単なお題を提示した．提示したお題は「桃太郎が犬・猿・雉を仲間にするシーン」とした．シナリオに登場するキャラクタを決めてもらい，本インタフェースを用いて各キャラクタのイメージに合う声を選択・登録してもらった．必要最低限の登場人物の登録が終わったらシナリオ執筆を開始してもらい，執筆したシナリオは必ず音声合成で読み上げて確認をしてもらった．シナリオの執筆には 20 分前後の時間を設定した．最後に，シナリオの執筆が終わった後に聞き取り調査を行った．聞き取り調査では，以下の 4 点について質問を行った．話者特徴量選択手法については，被験者 K にはニューラルネットワークの手法を用いてもらい，他の被験者にはユークリッド距離の手法を用いてもらった．

(1) 話者性を選択するインタフェースを用いることでイメージ通りの声を合成できたか？
(2) イメージに近い声で読み上げを行うことでセリフを修正しようと思ったか？
(3) イメージに近い声で読み上げを行うことでシナリオの書きやすさを感じたか？
(4) 普段の執筆作業と何か違いを感じたか？（執筆経験なしの被験者の場合，普段文章を書く時と何か違いはあるか？）

4.3 実験の結果と考察

表 2 に執筆実験後に各被験者に対して行った聞き取り調査の結果を示す．表中の各番号は前述の質問と対応する．

まず，キャラクタの個性のブレやセリフが自然でないと気づいてセリフの修正をすることは特になかった．この理由として，書いてもらったシナリオが 20 分程で書いた短いシナリオであったため，キャラクタの個性のブレが生じることや不自然なセリフを書くこと自体が起こらなかったためだと推測される．

次に，各被験者に注目して考察を行う．1 人目の被験者 K は漫画研究会に所属し，漫画の執筆経験がある．書きたいシナリオはなかったため，お題（桃太郎）を提示して執筆をしてもらった．

被験者 K に対する聞き取り調査から得られた意見として声質のレパートリーが少なく声の選択が上手くできなかったという点が挙げられた．これは，ニューラルネットワークによって 2 次元座標から話者特徴量を復元する際に 512 次元の特徴量を復元しきれないためだと考えられる．4.1 節で記述した類似性評価では，座標から復元した話者特徴量で合成した音声は最も低い評価となっていた．現状のニューラルネットワークによるモデルの品質では，音声合成によるシナリオ執筆への影響を評価することは難しいと考えられる．また，話速や読み方やアクセントなどの読み上げ方が物足りないという意見も得られた．この理由としては，被験者 K がコメディを主体としたシナリオを書くことが多いため，読み上げの話速をもっと早くしたいと考えることがあったためだと考えられる．しかし，本システムの音声はあくまで執筆作業を支援するためのものであり，音声そのものを読み手に提供するものではないため，個々の音声の編集というライターへの余計な作業を不要としている．読み上げの際の話速やアクセントなどを自動で設定する機能については今後の課題である．

次に，2 人目の被験者 W から得られた結果について考察する．被験者 W はゲームのシナリオを執筆した経験があり，書きたいシナリオがあったため，お題は提示せず実験を行った．被験者 W の実験では，被験者 K に対する実験で指摘された声のレパートリーの問題を解決するためユークリッド距離を用いた近傍話者音声探索の手法によって音声を合成した．

聞き取り調査の結果によると，被験者 K に対する 1 回目の実験と比較して望んだ声を選べているが，老人の声に関しては合成することができなかった．これは，音声合成モデルの学習に使用したデータセットに老人の声が含まれていなかったためだと考えられる．学習データにない声の生成は今後の課題である．一方，女の子の声はイメージに近い声を選択することができていた．そして，イメージに近い声で読み上げを行うことでイメージ像が明確になりセリ

フを書きやすくなると感じたという回答が得られた．さらに，本システムはチーム間でのキャラクタイメージの共有に役立つのではないかと考えられるという意見が得られた．

執筆経験がない被験者 2 人にお題を提示して実験をしてもらった結果，両者ともイメージに近い声を生成できていたという回答が得られた．また，イメージに近い声で読み上げることでキャラクタ性を維持したままセリフが書けたという意見やキャラクタの設定がしやすくなったという意見が得られ，この結果から執筆経験がない者は音声を付与することでキャラクタの設定が明確化され執筆がしやすくなるのではないかと考えられる．

本実験の結果よりシナリオ執筆経験者と未経験者では，キャラクタのイメージの仕方に違いがあることが考えられる．イメージ通りの声を合成できたかという質問に対して経験者は老人の声を合成できなかったと答えているが，未経験者は合成に納得していた．未経験者の場合は元々のイメージが明確ではないため，本システムによって合成される音声によってイメージの構築が行われたのではないかと推測される．一方経験者は元からある程度形作られたイメージを持っているためか納得できない声があると感じたのだと推測される．この結果から本システムを用いることで未経験者に対してはキャラクタのイメージの構築を支援し，経験者に対しては既にあるイメージの具体化を助けることができるのではないかと考えられる．

5. おわりに

本研究では，二次元平面上で声の位置関係が視覚化され，ライターがキャラクタのイメージに近い音声を探せるインタフェースを開発した．また，そのインタフェースを搭載したシナリオ執筆支援システムによってシナリオを執筆してもらうことでライターの執筆活動にどのような影響があるかの評価を行った．

シナリオ執筆支援システムを用いた実験において，1 人目の被験者 K にはニューラルネットワークの手法を用いたが良い結果が得られなかった．今後は，2 次元座標から話者特徴量を復元する手法の改善や話速やアクセントを考慮した読み方の改善など検討したい．2 人目以降はユークリッド距離に基づく最近傍話者探索の手法を使って執筆実験を行った．その結果，被験者 K に対する実験よりも被験者のイメージに近い声を合成できていた．執筆に対する影響についても，イメージ通りの声で読み上げを行うことでセリフと自分の中のイメージ像が結びつきやすくなったという回答やシナリオチームでのイメージの共有に役立つという回答が見られ，音声合成によってキャラクタ設定の明確化やチーム内でのイメージの共有に役立つと考えられる．

本システムを用いた執筆実験ではセリフの修正を行うなどの執筆活動への影響は見られなかったが，イメージの構築や具体化を助けることで執筆を支援することができるのではないかと考えられる．今後は，キャラクタのイメージに注目した評価実験など実施していきたい．

参考文献

[1] ［GDC 2021］シナリオライターが「台本の読み合わせ」をする意義とは。現役声優によるキャラクター作成術入手先 ⟨https://www.4gamer.net/games/999/G999905/20210720032/⟩ (2023.12.11).
[2] VOICEROID ｜製品情報｜ AHS(AH-Software) 入手先 ⟨https://www.ah-soft.com/voiceroid/⟩ (2023.12.11).
[3] Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., et al ”Transfer learning from speaker verification to multispeaker text-to-speech synthesis.” Advances in neural information processing systems 31 (2018).
[4] 鈴木勘太, 杉本徹. ”Encoder-Decoder モデルを用いた文章表現を豊かにする執筆支援システム.” 日本感性工学会論文誌 21.2 (2022): 257-265.
[5] 戀津魁, 菅野太介, 三上浩司, 近藤邦雄, 金子満. ”映像制作支援のためのシナリオ記述・構造化システムの開発.” 芸術科学会論文誌 10.3 (2011): 129-139.
[6] 栗山秀平, 伊藤克亘. ”画像・音声データを用いた対話エージェントのキャラクター設計支援.” 第 84 回全国大会講演論文集 2022.1 (2022): 853-854.
[7] 酒井えりか, 伊藤彰教, 伊藤貴之. ”ゲームキャラクタと声質の傾向分析.” 画像電子学会研究会講演予稿 画像電子学会第 277 回研究会講演予稿. 一般社団法人 画像電子学会, 2016.
[8] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., et al ”Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.” 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018.
[9] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S. ”X-vectors: Robust dnn embeddings for speaker recognition.” 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018.
[10] Cooper, E., Lai, C. I., Yasuda, Y., Fang, F., Wang, X., Chen, N., Yamagishi, J. ”Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings.” ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
[11] Yamamoto, R., Song, E., Kim, J. M. ”Parallel Wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram.” ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
[12] Sonobe, R., Takamichi, S., Saruwatari, H. ”JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis.” arXiv preprint arXiv:1711.00354 (2017).
[13] Takamichi, S., Mitsui, K., Saito, Y., Koriyama, T., Tanji, N., Saruwatari, H. ”JVS corpus: free Japanese multi-speaker voice corpus.” arXiv preprint arXiv:1908.06248 (2019).
[14] x-vector extractor for Japanese speech 入手先 ⟨https://github.com/sarulab-speech/xvector_jtubespeech⟩ (2023.12.09).
[15] McInnes, L., Healy, J., Melville, J. ”Umap: Uniform manifold approximation and projection for dimension reduction.” arXiv preprint arXiv:1802.03426 (2018).