社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

VQVAEによって獲得されたキャラクター演技スタイルに基づく多話者オーディオブック音声合成

中田　亘†　郡山　知樹†　高道慎之介†　齋藤　佑樹†　井島　勇祐††
増村　亮††　猿渡　洋†

† 東京大学　〒113–8654 東京都文京区本郷 7-3-1
†† 日本電信電話株式会社　〒239–0847 神奈川県横須賀市光の丘 1-1
E-mail: †nakata-wataru855@g.ecc.u-tokyo.ac.jp

あらまし　本研究では，Vector Quantized Variational AutoEncoder (VQVAE) を用いたキャラクター演技スタイルの抽出，及びそれを用いた多話者オーディオブック音声合成を提案する．声優によるオーディオブック音声では，登場人物の属性などにより異なるキャラクター演技スタイルが含まれため，オーディオブック音声合成においても異なるキャラクター演技スタイルを実現することが望まれる．一方で，テキスト情報のみから登場人物の属性と対応するキャラクター演技スタイルを推測することは困難である．そこで本研究では，音声からキャラクター演技スタイルを抽出しそれに基づく多話者オーディオブック音声合成を提案する．主観評価では，提案法を用いることにより，より原音声に近いキャラクター演技スタイルが実現できることが確認された．
キーワード　オーディオブック音声合成，VQVAE，キャラクター演技スタイル

Multi-speaker Audiobook Speech Synthesis using Discrete Character Acting Styles Acquired by VQVAE

Wataru NAKATA†, Tomoki KORIYAMA†, Shinnosuke TAKAMICHI†, Yuki SAITO†,
Yusuke IJIMA††, Ryo MASUMURA††, and Hiroshi SARUWATARI†

† The University of Tokyo　7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
†† Nippon Telegraph and Telephone Corporation　Hikarinooka 1-1, Yokosuka-shi, Kanagawa 239-0847, Japan.
E-mail: †nakata-wataru855@g.ecc.u-tokyo.ac.jp

Abstract　In this paper, we propose a method of extracting discrete character acting styles using vector quantized variational autoencoder (VQVAE) and multi-speaker audiobook speech synthesis based on extracted character acting styles. In audiobook corpora uttered by voice talents, the speech utterances contain acting depending on the character's attributes. Such acting should also be contained in synthesized audiobooks. However, predicting proper acting style and character attributes is still a hard challenge. To this end, we propose a method for extracting character acting styles from audiobook speech and conditioning TTS models by the extracted character acting styles to synthesize speech with character acting. The subjective evaluation shows that the proposed method achieves a closer character acting style to the ground truth speech.
Key words　Audiobook speech synthesis, VQVAE, Character acting styles

1. はじめに

近年，テキストから対応する音声を合成する音声合成では，深層学習の発展により単純な読み上げにおいて人間の自然音声に匹敵する品質が可能になりつつあり [1]，より表現力豊かな音声合成の実現に向けた音声コーパスの整備 [2], [3]，音響モデリング [4], [5] の研究開発が進められている．特に本研究では，音声合成を用いたオーディオブック生成（オーディオブック音声合成 [6]）で有用な音響モデリング手法に焦点を当てる．オーディオブック制作には長時間に及ぶ音声収録が必要となり，多大な労力や資金が必要となる．そこで，音声の収録をせずともオーディオブックのコンテンツを生成できるオーディオブック音声合成が有用である．声優によるオーディオブック音声では，文脈から推測される感情やキャラクターに合わせて演技が変化

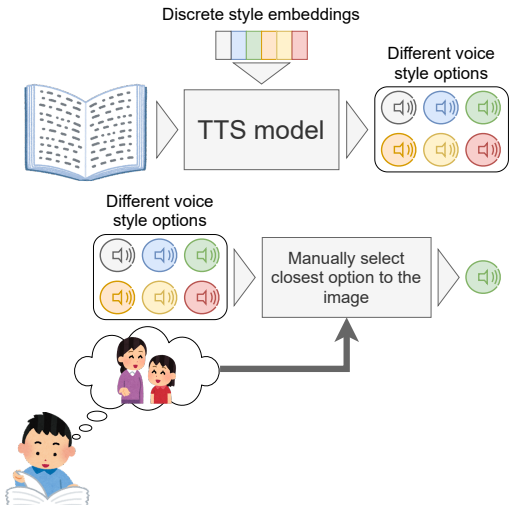


図 1: 本研究で目指す半自動オーディオブック音声合成
Fig. 1 Semi-automatic audiobook speech synthesis.

する [7]．これは，聴取者の作品の理解や興味を引きつける一助となっている．よってオーディオブック音声合成においても，キャラクターや，文脈に応じて適切な演技を実現することが望まれる．また，文学作品の解釈は個々人によって異なるため，オーディオブック音声合成においても，オーディオブックの作成を行うクリエイターの解釈を反映できるものが望ましい．

本研究では，音声の発話において言語情報と独立に現れ，F0，話速，パワーなどの韻律に主にその特徴が見られる，音声のスタイルに注目する．収録された音声から音声のスタイルを表す潜在空間へマッピング（本稿ではスタイルの獲得と呼ぶ）することで，任意のスタイルを実現するスタイル制御や，他の話者から獲得したスタイルを任意の話者の声で実現するスタイル転写などへの応用が可能になる．

音声のスタイルの獲得，制御は感情音声合成において研究されてきた [4] [8] [5]．感情音声合成では，同じテキストに対して複数の出力が存在する 1 対多の関係がある．例えば「食べすぎてつらい」というテキストでも，嬉しそうに読み上げる場合とつらそうに読み上げる場合とでは，音声が大きく異なる．このような 1 対多の関係があるため誤差最小化などの 1 対 1 の関係を仮定するモデルが不適切という問題がある．この問題に対処しながら，音声スタイルを制御する手法として教師なし学習によるスタイルの獲得 [4]，感情ラベリングが行われたコーパスの活用 [8] が提案されてきた．教師なし学習によるスタイルの獲得では変分オートエンコーダ（Variational AutoEnoder: VAE) を用いた手法 [5] が提案されている．この手法では，音声のスタイルを連続な潜在空間へ射影する．合成時には連続空間から任意の点を抽出して合成をすることが可能になるため，多様な音声スタイルを獲得することが可能になる．一方で潜在空間は人間による解釈が難しいため，オーディオブック音声合成に利用するためにクリエイターは（1）連続的な潜在空間からの任意点選択と（2）選択した潜在変数を用いて音声合成を行った結果の聴取を反復し，適切な音声スタイルが実現されているかどうかを確認する必要がある．このように，連続な潜在空間に音声スタイルをマッピングした場合，オーディオブック音声合成において多大な労力がかかることが考えられる．一方で感情ラベルが付与されたコーパスを用いた場合，音声スタイルのモデル化が感情ラベルに依存するため，多様な音声スタイルを実現することが難しい．

そこで本研究では，オーディオブック製作者が介在する半自動オーディオブック音声合成に取り組む（図 1）．本研究における半自動とは，入力テキストや，文脈などから推測可能な言語情報は TTS モデルが自動的に生成するものと仮定し，クリエイターの解釈により音声スタイルが変化する部分に関しては調整の余地を残し，その解釈を合成音声に反映できるようにする．特に本研究はキャラクター演技スタイルに焦点を当てる．キャラクター演技スタイルとは，音声スタイルの中でもキャラクターを声優が演技することにより実現される音声スタイルを指す．キャラクター演技スタイルは作品から想起されるキャラクターの属性（年齢，性別，性格など）により変化するため，クリエイターの解釈により音声スタイルが変化すると考えられる．提案法は，キャラクター演技スタイルは離散的であるという仮定を元に，話者により音声スタイルの異なる多話者オーディオブックコーパスを用いて学習され，Vector Quantized Variational Autoencoder (VQVAE) [9] により離散的なキャラクター演技スタイルを獲得，またそれに基づく多話者オーディオブック音声合成を可能とする．キャラクター演技スタイルを抽出するために，提案法では話者認証に用いられる ResCNN [10] を用いることにより時系列上で不変な音声のスタイルを抽出する．また，VQVAE を用いることによりオーディオブック合成を行うクリエイターは各文のキャラクター演技スタイルを離散的な選択肢から選択することが可能となるので連続的な潜在空間から選択するより簡便に合成することが可能になる．具体的にクリエイターは（1）有限個の離散表現から合成された音声サンプルを聴取（2）所望の音声スタイルが実現されているサンプルを選択することにより，オーディオブックを合成する事が可能になる．この手順では，繰り返しが存在しないため，大幅なオーディオブック制作の簡便化が期待される．加えて ResCNN の出力に対して，speaker adversarial classification [11] を行うことにより，キャラクター演技スタイルを話者非依存にする．これは，離散表現の選択によって合成音声の話者性が変化してしまうことを避け，キャラクター演技スタイルのみを制御する狙いがある．

実験では先行研究で提案された，文脈を考慮した音声合成を実現する音声合成モデルをベースラインとして比較を行う．評価では，合成音声の自然性，話者類似性，多様性，話者間のキャラクター演技スタイルの転写の 4 つの観点から評価を行う．結果から，音声の自然性，話者類似度を大きく損なう事なく，多様なキャラクター演技スタイルが実現できることを示す．また，キャラクター演技スタイルの制御，話者間の転写の可能性を示唆する．また，本研究は近年新たに構築された多話者オーディオブック音声コーパスである J-MAC [12] を用いた音声合成が可能であることを示す．

2. 提案する音声合成モデル

2.1 基本構造

図 2 に提案する音声合成モデルの構造を示す．提案するモデルは，FastSpeech2 [13] をベースとし，そこに BERT [14] から得られる文脈情報，VQVAE から得られるキャラクター演技スタイル及び話者分散表現により条件付けされる．これにより，学習時には発話レベルでキャラクター演技スタイルの獲得を行う．また，推論時には合成したい話者の ID と，VQVAE の学

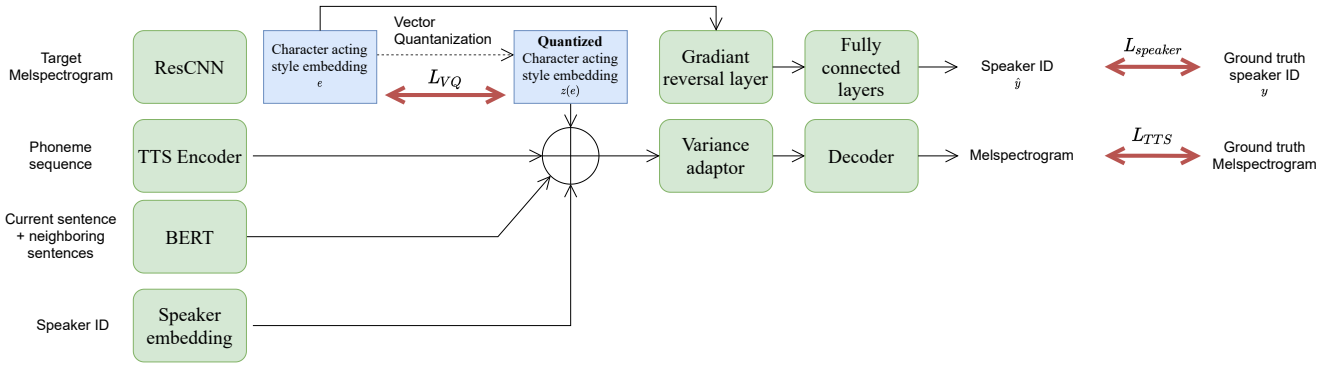


図 2: 提案法のモデル構造．Variance Adaptor が予測する F0, エネルギー，音素継続長に関する損失は省略されていることに注意されたい．

Fig. 2　The model architecture of proposed method.

習によって獲得されたコードブックから任意の発話スタイルを選択することにより，所望のキャラクター演技スタイルに対応した所望の話者の音声合成が可能となる．VQVAE によるキャラクター演技スタイルの抽出では，ResCNN [10] を用いる．また，ResCNN の出力に対して speaker adversarial classification を行い，ResCNN により得られる分散表現が話者非依存となるように学習を行う．損失関数を以下のように定義する．

$$L = L_{\text{TTS}} + \lambda_1 L_{\text{VQ}} + \lambda_2 L_{\text{speaker}} \tag{1}$$

$$L_{\text{VQ}} = \| \text{sg}[z_e(x)] - e\|_2^2 + \beta \| z_e(x) - \text{sg}[e]\|_2^2 \tag{2}$$

$$L_{\text{speaker}} = \mathit{CrossEntropyLoss}(y, \hat{y}) \tag{3}$$

ここで，L_{TTS} は音声の再構成誤差に関する損失，L_{VQ} は VQ-VAE のコードブックの学習に関する損失，L_{speaker} は speaker adversarial classification に関する損失である．また λ_1 及び λ_2 は各損失のゲインを調節するためのハイパーパラメータである．また sg は勾配逆伝搬を止める操作を指している．$z_e(x)$ は入力 x に対する ResCNN 出力，e は $z_e(x)$ に L2 距離が最も近いコードブックを指す．また，β は，ResCNN 出力がコードブックから離れるような更新を抑制するためのハイパーパラメータである．L_{speaker} に関する損失は，原音声の話者 y と推定された話者 $\hat{y}$ の予測分布から得られる交差エントロピーロスから定義した．本研究では, [9] とは異なり，発話レベルの潜在ベクトルを抽出していることに注意されたい．

2.2　ResCNN に基づくキャラクター演技スタイルの抽出

音声のスタイルを条件付ける音声合成では通常，Reference Encoder [15] が使用される [4], [5]．これは，Reference Encoder が，再帰型ニューラルネットワークを内包するため，発話内で変化する韻律要素（例えば強調）を抽出できるためである．一方で，本研究で用いる ResCNN は，話者認証で主に用いられるモデルであり，発話内で不変な要素（すなわち話者性）を抽出できる [16]．本研究で扱うキャラクター演技スタイルは発話内で不変な要素であると見做されるため，ResCNN の使用が好ましい．ただし，多話者コーパスを用いて ResCNN を学習する場合，キャラクター演技スタイルに加えて，同じ発話内で不変である話者性も抽出されてしまう恐れがある．そのため本研究では，ResCNN から抽出される特徴量に対して speaker adversarial classification [11] を行うことで，その特徴量を話者非依存にする．これにより，ResCNN から抽出される特徴量とそれを離散化したコードブックが，キャラクター演技スタイルのみを表現すると期待される．

3.　実　　験

実験では提案する音声合成モデルを，テキストと話者情報のみを入力とする TTS モデルと比較した．ベースラインの TTS モデルには周辺の複数文から得られる言語特徴を入力に用いるモデル (MultiSentences) [3] を用いた．MultiSentences と提案法の相違点は，キャラクター演技スタイルによる条件付けのみであり，それ以外の点では同一である．すなわち両者の違いは，図 2 における ResCNN，ベクトル量子化，speaker adversarial classification の有無である．また，元論文では Tacotron2 [1] をベースに MultiSentences を実装しているが，本研究では FastSpeech2 [13] をベースとして実装した．加えて単語レベルの文脈情報ではなく，文レベルの文脈情報による条件付けを行った．実装は Github 上で第一著者が公開している FastSpeech2 の日本語実装[1]をベースとした．

3.1　実 験 条 件

データセットには J-MAC [12] 及び J-KAC [3] を使用した．J-MAC は多話者日本語オーディオブック音声（39 話者，31.5 時間）であり，J-KAC は単一話者日本語オーディオブック音声（1 話者，9 時間）である．全体の音声の長さは約 40.5 時間である．全ての音声サンプルは事前に文レベルに分割した後，Julius [17] を用いて音素アライメントを取得し，22.05 kHz にダウンサンプリングした．train/valid/test セットはそれぞれ 20265/97/81 文とした．test セットは 1 つの作品からなり，train/valid セットと test セットの間で作品の重なりは無い．メルスペクトログラムは，フレーム長を 1024 サンプル，フレームシフトを 256 サンプル，次元数を 80 次元として生成した．また，図 2 の TTS Encoder 入力には音素の 256 次元の one-hot ベクトルとアクセントの 256 次元ベクトルの和を用いた．FastSpeech2 の Variance Adaptor は音声の F0，エネルギー，音素継続長を予測するように構成した．

BERT の事前学習には，日本語 Wikipedia を利用した．BERT のモデルサイズ設定は，Transformer 層数 $L = 2$, 隠れ層サイズ $H = 128$ とした．これは BERT-tiny [18] のモデル設定と同様である．音声合成モデルの訓練時には，BERT の word embedding のみ重みを固定した．BERT のトークンには subword を用いた．

(注1)：https://github.com/Wataru-Nakata/FastSpeech2-JSUT

最適化手法には，Adam [19] ($\beta_1 = 0.9, \beta_2 = 0.98$) を使用した．学習率に関しては先行研究 [20] と同様にスケジューリングを行った．ウォームアップステップ数は 4000 である．バッチサイズは 64 とした．ベースラインの損失関数は Variance Adaptor の各出力の L1 誤差及び Mel spectrogram 出力の平均平方誤差を使用した．提案法の損失関数は 2.1 節で説明した通りである．損失関数の重みを調整するハイパーパラメータは $\lambda_1 = 1$, $\lambda_2 = 0.1, \beta = 0.1$ とした．提案法の VQVAE に関してはコードブックサイズは 64，各ベクトルを 256 次元とし学習を行った．各モデルは 20 万ステップ学習を行った．

ボコーダーには HiFi-GAN [21] を使用した．HiFi-GAN の重みは公式実装[2]にて公開されている UNIVERSAL_V1 を使用した．HiFi-GAN の重みのファインチューニングは行わなかった．

3.2 評価指標

評価では自然性，話者類似性，話者間のキャラクター演技スタイルの転写，多様性の 4 つの観点から評価を行った．

合成音声の自然性に関しては，Mean Opinion Score (MOS) テストによる主観評価を行った．各評価者は合成音声の自然性を 5 段階（1：とても悪い，5：とても良い) の指標で評価した．MOS テストはクラウドソーシングによる主観評価システムで実施し，評価者数は 120 名であり，各評価者は 20 発話を評価した．キャラクター演技スタイルに関しては，各発話に対応する原音声から提案法を用いて抽出した．また，話者類似性に関しては GitHub 上で公開されている Resemblyzer[3]を用いて d-vector を抽出し，その結果を t-SNE を用いて次元削減を行い可視化し d-vector の分布を確認した．Resemblyzer の d-vector 抽出手法は [22] をベースとしている．

話者間のキャラクター演技スタイルの転写に関しては，話者 A から話者 B に転写した際に話者 A の原音声と同様のキャラクター演技スタイルが実現されているかどうかの XAB テストをクラウドソーシングによる主観評価にて実施した．評価者数は 60 名であり，各評価者は 10 発話を評価した．話者 A は J-KAC の話者，話者 B は訓練データに含まれる話者から J-KAC の話者及び J-MAC から男性話者 5 名，女性話者 4 名を選んだ．各評価者は「以下の音声は参照音声のキャラクター演技スタイルを元に他の話者にコピーしたものです．どちらがより参照音声のキャラクター演技スタイルに近いですか？」という問に対して，J-KAC の原音声とそれぞれ提案法とベースラインの合成音声のペアから選んだ．

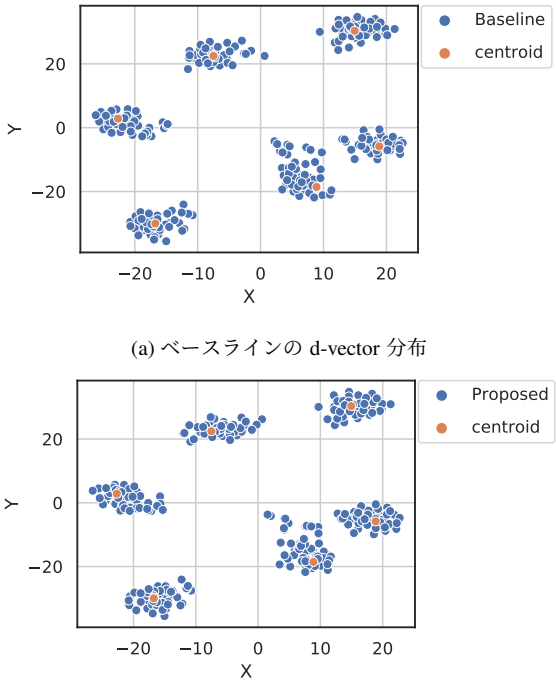
最後に多様性の評価に関しては，各コードブックで条件付けされた音声を合成した際に音声が変化している事を確認するために，異なるキャラクター演技スタイルで条件付された音声間の平均メルケプストラム歪 [23]（Mel-Cepstral Distortion: MCD）及び各合成音声の F0，パワー，話速の平均，標準偏差を可視化した．MCD の計算に際しては，各合成音声の系列長が異なるため，FastDTW [24] を用いて系列長を合わせた上で計算した．ピッチ，パワーに関しては，librosa[4]を用いて抽出し，話速に関しては，入力音素列長と対応する出力音声の長さから計算したそれぞれの特徴量に対し，各キャラクター演技スタイル，話者ごとに平均，標準偏差を求め，可視化を行った．MCD の計算

(注2)：https://github.com/jik876/hifi-gan
(注3)：https://github.com/resemble-ai/Resemblyzer
(注4)：https://librosa.org/

表 1: MOS による自然性の評価結果 ± 以降は結果の 95% 信頼区間を示す．

Table 1 The result for Naturalness MOS. The ± indicates 95% confidence intervals.

Method	MOS
Baseline	2.94 ± 0.039
Proposed	2.86 ± 0.040



(a) ベースラインの d-vector 分布

(b) 提案法の d-vector 分布

図 3: 各手法の d-vector 分布．centroid は各話者の原音声の重心を示す．

Fig. 3 d-vector distribution of compared methods. Orange circles indicate the centroids of d-vectors from ground truth speech samples for each speaker.

式は以下で与えられる．

$$\mathrm{MCD} = \frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\sum_{k=1}^{K}\left(c_{t,k} - \hat{c}_{t,k}\right)^2}$$

$c_{t,k}$ および $\hat{c}_{t,k}$ は，それぞれ原音声と合成音声の t フレーム目 k 次元目のメルケプストラム係数である．また，メルケプストラム係数の次数は $K = 59$ とした．

3.3 結　　果

3.3.1 合成音声の自然性

表 1 にベースラインモデルと提案法の自然性 MOS の評価結果を示す．音声の自然性に関しては，僅かではあるが提案法が有意に悪いという結果となった．これは，キャラクター演技スタイルが音声に付与されることにより，より表現豊かな合成音声（男性話者による女性キャラクター演技など）が生成され，裏声などの一部のキャラクター演技スタイルで合成音声が不安定になったためと考えられる．

3.3.2 話者類似性

図 3 に Resemblyzer を使用して抽出した d-vector の話者別，手法別の t-SNE プロットを示す．各手法で合成された音声の

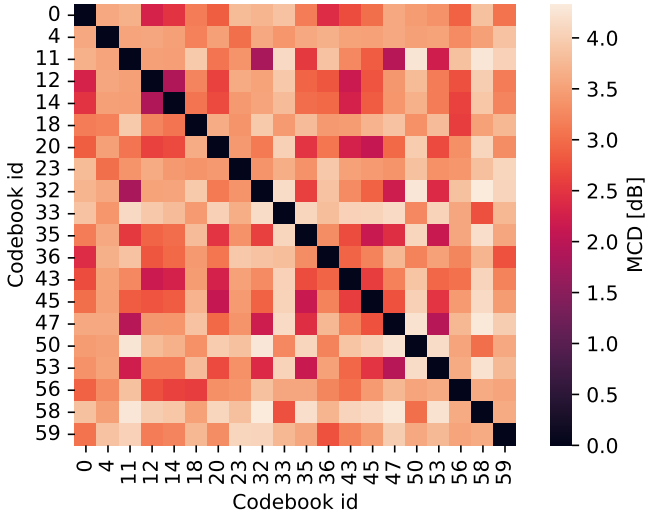


図 4: 各コードブック（codebook id) で条件付けされた合成音声間の MCD.

Fig. 4 MCD between synthesized speeches conditioned by different codes from VQVAE

d-vector は青い点で，各話者の原音声に対応する d-vector の重心はオレンジ色の点で示される．話者類似性に関して手法による分布の大きな差異は確認されなかった．これは，キャラクター潜在スタイルを用いた場合でも，話者性は大きく損なわれていない事を示唆している．

3.3.3 音声の多様性

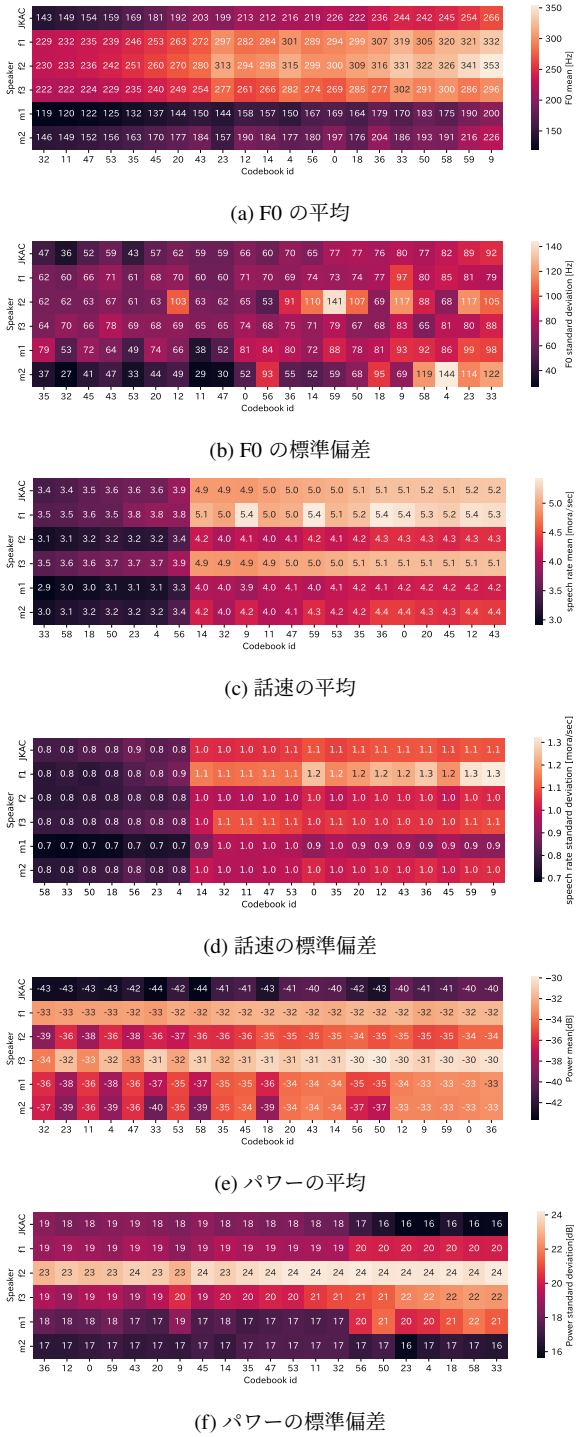
学習が収束するにつれて，VQVAE のコードブック（キャラクター演技スタイル）は一部でしか更新が行われなかった．よって多様性の評価に関しては，全ての学習に用いた音声サンプルから抽出したキャラクター演技スタイルで出現するもののみを用いて評価を行った．

図 4 に各コードブックで条件付けされた音声同士の MCD を示す．異なるコードブックで合成した音声間の MCD の最低値は約 1.7 [dB] であった．この事から，提案法では多様な音声スタイルを獲得できていると考えられる．

図 5 に各コードブック及び各話者により条件付けされた合成音声の F0，パワー及び話速の平均，標準偏差を示す．

F0 及び話速に関しては，コードブックを変化させることにより各話者で共通の変化が見られた．これは，コードブックの話者非依存性を示唆しており，speaker adversarial classification によりコードブックから話者情報を分離できていることを示唆している．同様に，標準偏差に関しても話者によらないコードブックによる変化が見られた．このことから，ある話者の演技音声から得られたコードブックを別の話者へ転写した際に，同様なキャラクター演技スタイルを実現するために重要な F0 に関して，相対的な音高の転写が可能であることが示唆された．また，F0，話速の平均に関して一番小さいもの（一番左の列）と，一番大きいもの（一番右の列）で 1.5 倍ほど値が異なることから，多様な音声スタイルが実現されている事が確認された．

一方でパワーに関しては各コードブックを変化させることによる各話者での共通の変化が見られなかった．このことから，提案法を用いてのパワーの制御は困難であると思われる．



(a) F0 の平均

(b) F0 の標準偏差

(c) 話速の平均

(d) 話速の標準偏差

(e) パワーの平均

(f) パワーの標準偏差

図 5: 話者ごと及び各コードブック (codebook id) で条件付けされた合成音声の F0，パワー，話速の平均及び標準偏差．JKAC は J-KAC コーパスの話者（男性），m1,m2 は男性話者，f1,f2,f3 は女性話者を示す．それぞれ，平均の数値が左から右にむけて昇順になるようにソートされている．

Fig. 5 F0, power and speech rate of synthesized speech by speaker and by learnt VQVAE codes (codebook id). JKAC indicates speaker of J-KAC corpus. m1, m2 and f1, f2, f3 indicates male and female speaker respectively. Each plot is sorted in ascending order from left to right.

3.3.4 話者間のキャラクター演技スタイルの転写

表 2 に話者 A から話者 B に転写した際に同様のキャラクター

表 2: 話者間のキャラクター演技スタイル転写に関する XAB テスト結果

Table 2 The Result for XAB test of character acting style transfer between speakers.

Baseline	Proposed	p-value
0.467	0.53278	0.022

演技スタイルが実現されているかどうかの XAB テスト結果を示す．提案法はベースラインに比べ僅かではあるものの有意に転写した後に参照音声と同様のキャラクター演技スタイルが実現されている事が示された．この事から，提案法ではキャラクター演技スタイルの転写が可能である事が確認された．

4. ま と め

本研究では，半自動オーディオブック音声合成において有用な，離散的なキャラクター演技スタイルを VQVAE を用いて獲得し，それを用いて多話者オーディオブック音声を合成する手法を提案した．評価実験により，キャラクター演技スタイルの条件付けにより自然性が僅かに劣化するものの，話者類似性ではベースラインと同様な性能を示した．また，話者間のキャラクター演技スタイルの転写では，獲得したキャラクター演技スタイルの話者非依存性を示し，多様性の評価では得られたキャラクター演技スタイルが多様な音声スタイルをマッピングしていることを示した．

一方で，評価方法に関しては課題が残る．本研究の目的は，半自動オーディオブック音声合成であるが，複数文のオーディオブック音声に対する評価は行われておらず，評価は全て文レベルで行われているため，提案法がオーディオブック音声として適切なのかの評価が行われていない．また，実際に製作者が所望の音声スタイルを実現できるのかが不透明である．このことから，今後の課題としてより適した評価方法の検討が挙げられる．

文 献

[1] J. Shen, R. Pang, R.J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R.A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, “Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4779–4783, 2018.

[2] 高道慎之介, 中田亘, 郡山知樹, 丹治尚子, 井島勇祐, 増村亮, 猿渡洋, “J-KAC: 日本語オーディオブック・紙芝居朗読音声コーパス,” 情報処理学会研究報告, 第 2021-SLP-137 巻, pp.1–4, 2021.

[3] N. Wataru, K. Tomoki, T. Shinnosuke, T. Naoko, I. Yusuke, M. Ryo, and S. Hiroshi, “Audiobook speech synthesis conditioned by cross-sentence context-aware word embeddings,” Proc. 11th ISCA Speech Synthesis Workshop (SSW 11), pp.211–215, 2021.

[4] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai, “End-to-end emotional speech synthesis using style tokens and semi-supervised training,” 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp.623–627, 2019.

[5] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, “Learning latent representations for style control and transfer in end-to-end speech synthesis,” ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6945–6949, 2019.

[6] J. Pan, L. Wu, X. Yin, P. Wu, C. Xu, and Z. Ma, “A chapter-wise understanding system for text-to-speech in chinese novels,” ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)IEEE, pp.6069–6073 2021.

[7] É. Székely, J.P. Cabral, M. Abou-Zleikha, P. Cahill, and J. Carson-Berndsen, “Evaluating expressive speech synthesis from audiobook corpora for conversational phrases,” Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12), pp.3335–3339, May 2012.

[8] M. Kim, S.J. Cheon, B.J. Choi, J.J. Kim, and N.S. Kim, “Expressive text-to-speech using style tag,” Proc. Interspeech 2021, pp.4663–4667, 2021.

[9] A. van denOord, O. Vinyals, and k. kavukcuoglu, “Neural Discrete Representation Learning,” Advances in Neural Information Processing Systems (NIPS), vol.30, pp.••–••, 2017.

[10] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, “Deep Speaker: an end-to-end neural speaker embedding system,” CoRR, vol.abs/1705.02304, pp.••–••, 2017.

[11] Z. Meng, J. Li, Z. Chen, Y. Zhao, V. Mazalov, Y. Gong, and B.-H. Juang, “Speaker-invariant training via adversarial learning,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5969–5973, 2018.

[12] S. Takamichi, N. Tanji, and H. Saruwatari, “J-MAC corpus,” https://sites.google.com/site/shinnosuketakamichi/research-topics/j-mac_corpus.

[13] 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,” International Conference on Learning Representations, pp.••–••, 2021.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol.1, pp.••–••, June 2019.

[15] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R.A. Saurous, “Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron,” Proceedings of the 35th International Conference on Machine Learning, vol.80, pp.4693–4702, 2018.

[16] M. Chen, M. Chen, S. Liang, J. Ma, L. Chen, S. Wang, and J. Xiao, “Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding.,” Interspeech, pp.2105–2109, 2019.

[17] A. Lee and T. Kawahara, “Recent development of open-source speech recognition engine julius,” Proceedings : APSIPA ASC 2009 : Asia-Pacific Signal and Information Processing Association (APSIPA), pp.131–137, oct 2009.

[18] I. Turc, M. Chang, K. Lee, and K. Toutanova, “Well-read students learn better: The impact of student initialization on knowledge distillation,” arXiv, vol.abs/1908.08962, pp.••–••, 2019.

[19] D.P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 3rd International Conference on Learning Representations, ICLR 2015,, eds. by Y. Bengio and Y. LeCun, pp.••–••, 2015.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in Neural Information Processing Systems, vol.30, pp.••–••, 2017.

[21] J. Su, Z. Jin, and A. Finkelstein, “HiFi-GAN: high-fidelity denoising and dereverberation based on speech deep features in adversarial networks,” Interspeech 2020, pp.4506–4510, 2020.

[22] L. Wan, Q. Wang, A. Papir, and I.L. Moreno, “Generalized end-to-end loss for speaker verification,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4879–4883, 2018.

[23] R. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, vol.1, pp.125–128, 1993.

[24] S. Salvador and P. Chan, “FastDTW: Toward accurate dynamic time warping in linear time and space,” KDD workshop on mining temporal and sequential dataCiteseer, pp.••–••2004.