キャラクタ分散表現を用いた演じ分けを実現するオーディオブック音声合成*

☆中田 亘 (東大 工学部), 郡山 知樹, 高道 慎之介, 齋藤 佑樹（東大院・情報理工）
井島 勇祐, 増村 亮（NTT）, 猿渡 洋（東大院・情報理工）

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目覚ましい発展を遂げている．特に単純な読み上げにおいては，人間の自然音声に匹敵する品質の音声が合成可能になりつつあり [1]，今後の研究は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]，音響モデリング [3, 4] に焦点が当てられている．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 [5]）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，それに伴い，声優の長時間な拘束や，スタジオの使用料などの，多大な労力や資金が必要となる．そこで，音声収録を合成音声に置き換え手間を省く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に合わせて演技を行い声のスタイルが変化する [6]．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，文脈から推測される感情や，キャラクタに合わせた声質の変化が望まれる．

先行研究 [7] では，演技音声から，推測されるキャラクタの属性（性別，性格，年齢など）をクラウドソーシングを用いて収集し，音声からキャラクタ性を抽出している．しかしながら，本先行研究は離散的な属性を使用しており，プロの声優による多様な演技音声を反映できているとは考えにくい．また，この研究は，分散表現の抽出にとどまっており，実際に得られた分散表現に基づいた音声合成が可能かが示されていない．また我々はこれまでに VQVAE を用いたキャラクタ演技スタイルの獲得およびその制御を提案している [8]．この手法は人間によるキャラクタ演技スタイルの選択に依存しており，合成する文学作品に含まれるすべての発話に対して，キャラクタ演技スタイルの選択を行う必要がある．

本研究では，多様な演技の中でもキャラクタ演技に着目し，作品と作品内のキャラ発話の情報からキャラクタ分散表現を抽出し，キャラクタ性を内包する音声合成を実現する音声合成モデルを提案する．提案法は，各キャラクタ発話に対応するキャラクタ名，

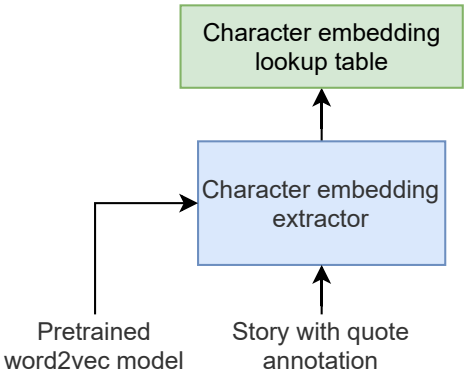

Fig. 1: Character embedding（Skip Gram）によるキャラクタ分散表現の獲得

および各キャラクタの発話文から適切な演技スタイルを推測し，音声を合成する．提案法は，まず，文学作品に含まれる言語特徴量からキャラクタを表す分散表現を獲得する．キャラクタ分散表現の獲得では，各キャラクタ発話に対する人間によるキャラクタ名のアノテーションを利用する．その後，音響モデルを獲得した分散表現において条件付けし，学習を行う．これによりオーディオブック音声制作の大幅な簡便化が望まれる．実験の結果から，提案法は従来法に比べ有意にキャラクタの演じ分けができることを示す．

2 提案する音声合成モデル

2.1 キャラクタ分散表現の抽出

Figure 1に，本研究で使用した Character embedding (Skip Gram)[9] によるキャラクタ分散表現の獲得方法の概略図を示す．本研究では，キャラクタ分散表現の抽出を目的として，映画のキャラクタ表現を抽出するのに有効であるとされている Character Embedding (Skip Gram) を使用した．この手法では，キャラクタ間の関係性を内包する特徴量を抽出するために，キャラクタの対話相手，およびその会話の内容などに基づき，各キャラクタの名前および会話内容を考慮した分散表現を獲得する．分散表現の獲得は，学習済み Word2vec のキャラクタ名に対応する分散表現に対して fine-tuning を行う形で獲得される．得られたキャラクタ分散表現のキャラクタ間コサイン類似度と人間により付与されたキャラクタ関連

*Audiobook Speech Synthesis based on Character embedding for Distinguishable Character Acting by NAKATA, Wataru, KORIYAMA, Tomoki, TAKAMICHI, Shinnosuke, SAITO, Yuki (The University of Tokyo), IJIMA, Yusuke, MASUMURA, Ryo (NTT), SARUWATARI, Hiroshi (The University of Tokyo)

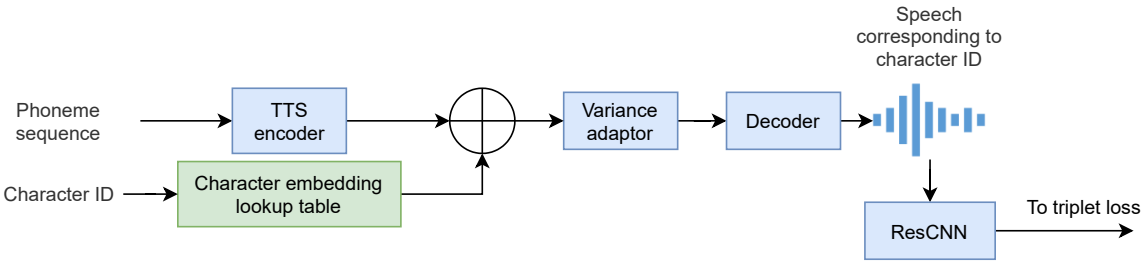

Fig. 2: 提案する音声合成モデルの構造

度 (Character relatedness) の間には高い相関があると報告されている．こうした，キャラクタの関連度によりオーディオブック音声合成においても有用であると考えられる．例えば，関連度の高いキャラクタ間では，親しみのこもった韻律が実現されていると考えられる．一方で主人公との関連度が低いキャラクタに関しては親しみのこもってない韻律が期待される．また，この分散表現抽出手法では，発話内容が考慮されるため，語尾の口調や一人称などから性別や性格などの情報が推測できると考えられる．

2.2 キャラクタ分散表現の音声合成への適用

Figure 2に本研究で提案する音声合成モデルの構造を示す．本研究で提案する音声合成モデルは，音響モデルとして FastSpeech2[10] をベースとし，それに2.1節で説明するキャラクタ分散表現で条件付けを行う．これにより，各キャラクタの名前，およびセリフを考慮した音声合成が期待される．また，キャラクタごとに異なる声質をより明示的に実現するために，学習時には ResCNN[11] を用いて合成音声の発話内で時間変化しない共通の特徴量を抽出し，それに対して triplet loss[12] を使用する．triplet loss は，感情識別などで使用されており [13]，その有効性が示されている．本研究では，anchor と同じキャラクタによる発話を positive，同一作品内の異なるキャラクタの発話を negative とした．これにより，同じキャラクタによる発話は特徴量空間で近く，違うキャラクタによる発話は特徴量空間で遠くなり，よりキャラクタ演技が明確になることが期待される．anchor,positive,negative サンプルのサンプリングは無作為に行う．損失関数は以下のように定義する．

$$L = L_{\mathrm{TTS}} + \lambda L_{\mathrm{triplet}} \tag{1}$$

$$L_{\mathrm{triplet}} = \max\{d(x, x_p) - d(x, x_n) + \mathrm{m}, 0\} \tag{2}$$

ここで，L_{TTS} は音声の生成誤差に関する損失であり，$d(\cdot)$ は二つの入力間の L2 ノルムを出力する関数，x は anchor サンプルから抽出した特徴量，x_p,x_n がそれぞれ positive サンプル（anchor と同じキャラクタ），negative サンプル (anchor と異なるキャラクタ）から抽出した特徴量である．また m は，positive サンプルと negative サンプルの距離をどれだけ離すかを定義するハイパーパラメータであり，λ は，L_{triplet} のゲインを調整するハイパーパラメータである．

3 実験

実験では以下に示す 3 つのモデルを比較した．

1. FS: 通常の FastSpeech2 キャラクタ分散表現**なし**
2. w/o triplet: 提案法 triplet loss **なし**
3. w/ triplet: 提案法 triplet loss **あり**

FS はキャラクタ分散表現を用いずにテキストと音声のみから学習した FastSpeech2 である．w/o triplet は，キャラクタ分散表現で条件付けした FastSpeech2 である．w/ triplet は，キャラクタ分散表現で条件付けした上で，triplet loss を用いて学習を行った FastSpeech2 である．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-KAC[2] を使用した．J-KAC は単一話者日本語オーディオブック音声（男性話者 1 名，9 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[14] 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/dev/test セットはそれぞれ 5129/100/81 発話とした．なお，test セットは 1 つの紙芝居作品からなり，train/valid セットと test セット間で作品の重なりは無い．よって，test セットに現れるキャラクタは学習データに存在しない．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 [15] を用いた．FastSpeech2 の Variance Adaptor は音声の F0，パワー，音素継続長を予測するように構成した．

最適化手法には，Adam[16] ($\beta_1 = 0.9$, $\beta_2 = 0.98$) を使用した．学習率に関しては先行研究 [17] 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

[1] https://github.com/Wataru-Nakata/FastSpeech2-JSUT/

プ数は 4000 とした．バッチサイズは 64 とした．損失関数 L_{TTS} は Variance Adaptor の各出力の L1 誤差及びメルスペクトログラム出力の平均平方誤差の和として定義した．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[18] 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．

Character Embedding (Skip Gram) の学習では，Gensim[19] の word2vec の実装を元に，BCCWJ コーパス[3] に含まれる文の中から NDC 区分表において 90 から始まる文学作品を抽出した上で word2vec の学習を行った．その後，2.1節で説明したキャラクタ分散表現の抽出を J-KAC に対して行った．キャラクタ分散表現の抽出は train/dev/test セットすべてに対して行ったが，test セットのキャラクタ分散表現が train セットのキャラクタ分散表現に影響することはない．キャラクタ分散表現の抽出に必要となる J-KAC に含まれる文学作品に対するキャラクタアノテーションデータに関しては，各発話に対して人間によるアノテーションを実施した.[4] また，w/ triplet のハイパーパラメータは $\lambda = 0.1$, $\mathrm{m} = 1.0$ とした．抽出に際して，非キャラクタ発話である地の文に関しては，学習データから取り除き，キャラクタ分散表現の獲得を行った．

3.2 評価指標

評価では，合成音声の自然性に加えキャラクタの演じ分けについて評価を行った．

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．各評価者は一発話に対応する合成音声を聴取し，その自然性を 5 段階の指標（1: とても悪い，5: とても良い）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 50 人で，各評価者は 18 発話を評価した．

キャラクタの演技分けについては，AB テストおよびテキスト回答による評価を行った．AB テストでは，文レベルで合成した音声を J-KAC で定義されている chapter レベル（2-5 発話）で連結したものとそれに対応するキャラクタアノテーションされたテキストを用意した．評価者は表示されたテキストを読んだ上で,「A,B 二つの音声を聴いて，下の台本をもとに，より演技分けが適切なほうを選んでください」という問に対して回答した．なお，今回の test セットには紙芝居音声が使用され，J-KAC で定義されている chapter レベル分割は，紙芝居一枚に対応する音声となっている．評価者数は 100 人で，各評価者は 12 個の設問に答えた．

一方でテキスト回答による評価では，AB テストと同じ設問に対して，その回答理由を記述し，どういった理由で A/B を選択したかの回答を集めた．評価者数は 50 人で各評価者は 6 問の設問に答えた．

[2] https://github.com/jik876/hifi-gan

[3] https://ccd.ninjal.ac.jp/bccwj/index.html

[4] アノテーションしたデータは NII のコーパスページにて公開されている．https://research.nii.ac.jp/src/J-KAC.html

Table 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.

Method	MOS
FS	3.08 ± 0.101
w/o triplet	2.99 ± 0.110
w/ triplet	2.91 ± 0.110

Table 2: AB テストの結果. 評価者はより演技分けが適切なほうを台本を元に選択した． **太字**は比較対象に対して有意な改善が得られたモデルを示している.

Method A	Scores	p-value	Method B
FS	0.478 vs. 0.523	0.396	w/o triplet
FS	0.453 vs. **0.548**	0.007	w/ triplet
w/o triplet	0.485 vs. 0.515	0.203	w/ triplet

3.3 結果

3.4 合成音声の自然性

Table 1に各手法の自然性 MOS 評価結果を示す．音声の自然性に関しては，w/ triplet と，FS を比較すると有意に劣化していることが確認された．これは，キャラクタ演技スタイルが音声が付与されることにより，合成音声の韻律が不安定になったためと考えられる．

3.5 キャラクタの演技分け

Table 2に AB テストによるキャラクタの演技分け評価結果を示す．AB テストによるキャラクタの演技分け評価では，FS と w/triplet 間でのみ有意差が確認された．このことから，キャラクタの演技分けを実現するためには，キャラクタ分散表現で TTS モデルを条件付けするだけでは，不十分であり，triplet loss を用いた学習が必要であると考えられる．

テキスト回答による評価によって得られた結果は，Universal Sentence Encoder (USE)[20] を用いて特徴量を抽出した後に，k 平均法を用いて 5 個のクラスタへと分割した．USE の重みには USE の著者らが公開している重み[5] を使用した．各クラスタごとの回答の傾向を Table 3に示す．各クラスタにおいて，適切だと選択された手法に偏りがある事がわかる．そこで，各クラスタに含まれる回答から，キーワード抽出を pke[21] を用いて行う事により手法間の差異を明らか

[5] https://tfhub.dev/google/universal-sentence-encoder/3

Table 3: クラスタごとに選択された手法

Cluster id	w/o CE	w/ CE	w/ CE w/ triplet
0	26	25	22
1	18	27	25
2	6	8	15
3	6	22	15
4	20	35	30

Table 4: pke を用いて抽出した各クラスタのテキスト回答に含まれるキーワード

Cluster id	Keywords
0	雰囲気, ナレーション, うまく 感情, 良かっ, うまく
1	演技 分け, おばけ, ギャップ, セリフ, 完成度
2	演技 わけ, 良かっ, キャラクター, 女の子, 演技 分け
3	女の子, セリフ, ナレーション, トーン, 棒読み
4	ナレーション, セリフ, イントネーション, 女の子, ありゃ

にした．Table 4に各クラスタから抽出したキーワードを示す．w/ triplet の評価が良かった cluster id 2 では,「演技分け」などのキーワードが多く含まれている．このことから，w/ triplet は，演技分けという観点において特に有用であると考えられる．一方で，w/ triplet の評価が相対的に低い cluster id 0,3,4 では,「ナレーション」というキーワードが見られた．これは，ナレーションに着目した際に，w/ triplet が選好されていないことを示している．このことから，w/ triplet では，ナレーションの部分において適切な音声が実現できていないことが示唆される．

4 まとめ

本研究では，オーディオブック作品内のキャラクタ発話に対応するキャラクタ名，および各キャラクタによる発話文を用いて，キャラクタ分散表現を抽出し，それに基づくキャラクタの演技分けを実現するオーディオブック音声合成を提案した．結果から，キャラクタ分散表現で音声合成モデルを条件付けするだけでは，不十分であり，triplet loss を用いて，キャラごとの声質を明示的に異なるものとすることが必要であることが確認された．今後の課題としては，ナレーションにおいても適切な音声スタイルを実現する音響モデリングがあげられる．

参考文献

[1] J. Shen et al., “Natural TTS synthesis by conditioning Wavenet on mel spectrogram predictions,” in *Proc. ICASSP 2018*, 2018, pp. 4779–4783.

[2] N. Wataru et al., “Audiobook speech synthesis conditioned by cross-sentence context-aware word embeddings,” in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 211–215.

[3] P. Wu et al., “End-to-end emotional speech synthesis using style tokens and semi-supervised training,” in *Proc. APSIPA ASC 2019*, 2019, pp. 623–627.

[4] Y.-J. Zhang et al., “Learning latent representations for style control and transfer in end-to-end speech synthesis,” in *Proc. ICASSP 2019*, 2019, pp. 6945–6949.

[5] J. Pan et al., “A chapter-wise understanding system for text-to-speech in Chinese novels,” in *Proc. ICASSP 2021*, 2021, pp. 6069–6073.

[6] É. Székely et al., “Evaluating expressive speech synthesis from audiobook corpora for conversational phrases,” in *Proc. LREC 12*, 2012, pp. 3335–3339.

[7] E. Greene et al., “Predicting character-appropriate voices for a TTS-based storyteller system,” in *Proc. INTERSPEECH 2012*, vol. 3, 2012, pp. 2207–2210.

[8] 中田亘 他, “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,” in **研究報告音声言語情報処理** *(SLP)*, vol. 2021, no. 23, 2021, pp. 1–6.

[9] M. Azab et al., “Representing movie characters in dialogues,” in *Proceedings of the 23rd CoNLL*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 99–109.

[10] Y. Ren et al., “FastSpeech 2: Fast and high-quality end-to-end text to speech,” in *International Conference on Learning Representations*, 2021.

[11] C. Li et al., “Deep Speaker: an end-to-end neural speaker embedding system,” *CoRR*, vol. abs/1705.02304, 2017.

[12] J. Wang et al., “Learning fine-grained image similarity with deep ranking,” *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1386–1393, 2014.

[13] J. Huang et al., “Speech emotion recognition from variable-length inputs with triplet loss function,” in *Proc. Interspeech 2018*, 2018, pp. 3673–3677.

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

[15] 藤井一貴 他, “韻律情報で条件付けされた非自己回帰型 End-to-End 日本語音声合成の検討,” in **情報処理学会研究報告**, vol. 2021-SLP-138, no. 16, 2021, pp. 1–6.

[16] D. P. Kingma, J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015,*, 2015.

[17] A. Vaswani et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[18] J. Su et al., “HiFi-GAN: high-fidelity denoising and dereverberation based on speech deep features in adversarial networks,” in *Interspeech 2020*, 2020, pp. 4506–4510.

[19] R. Řehůřek, P. Sojka, “Software framework for topic modelling with large corpora,” in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50.

[20] D. Cer et al., “Universal sentence encoder,” *CoRR*, vol. abs/1803.11175, 2018.

[21] F. Boudin, “pke: an open source python-based keyphrase extraction toolkit,” in *Proceedings of COLING 2016*, Osaka, Japan, 2016, pp. 69–73.