アクセント潜在変数の予測と制御が可能な TTS モデルによる方言音声合成の検討 *

☆山内一輝, 齋藤佑樹, 猿渡洋 (東大院・情報理工)

1　はじめに

テキスト音声合成 (Text-to-Speech: TTS) とは, 任意のテキストから対応する自然な読み上げ音声を合成する技術である. TTS は, あるテキストからパラ言語情報や非言語情報などに由来する多様な音声を合成するという One-to-Many mapping 問題であり, 自然な韻律の予測は重要かつ困難な問題となっている.

本研究の対象となる日本語は, ピッチの高低によってアクセントを表現するピッチアクセント言語であり, 同音異義語の弁別や方言音声としての知覚において, 合成音声アクセントの正確なモデリングは重要な役割を持つ. しかし, 現状の方言 TTS では, 1) 話者数の限られた方言におけるアクセント辞書の欠如, 2)TTS モデルの学習に利用できる十分な品質の方言音声収録の困難性といった課題に対処する必要がある.

そこで本研究では, 音声データが少数しか用意できずアクセント辞書も存在しないような方言を含めた, ありとあらゆる地域の方言アクセントを再現可能な TTS モデルを構築する. 提案モデルは, アクセント潜在変数 (Accent Latent Variable: ALV) をテキストから予測する ALV Predictor および ALV を参照音声から自動抽出する Reference Encoder を備えており, ユーザからの音声入力もしくは ALV 訂正入力に基づき, 合成音声のアクセントに誤りが生じた場合に, ユーザからのフィードバックに基づいて誤りを訂正するための教師信号を獲得できる. 実験的評価により, 望ましい ALV を与えて生成した音声が, アクセント情報を用いずに生成した音声よりも, 自然性および方言らしさに関する MOS スコアが高いことを示す.

2　関連研究

Yufune らはアクセント辞書が存在しない方言を想定し, Vector Quantised-Variational AutoEncoder (VQ-VAE) [1] を用いて音声データからアクセント情報 (ALV) を抽出する手法を提案した [2]. 一方で, テキストからアクセント情報を予測するモデルを作成するためには, 十分な語彙を含む学習データが必要となる. しかし, 現状存在する方言音声コーパスのサイズは非常に小さく, 十分な量の語彙を集めることができないという問題がある.

また, Fujii らは合成音声のアクセント誤りをユーザーが簡単に訂正することが可能な人間参加型音声合成の枠組みを提案した [3]. 韻律が制御可能な TTS モデルを用いることで, アクセント辞書が存在しないような方言音声合成に対しても音声の合成時にユーザーが適切なアクセント情報を与えることができる. しかし, Fujii らのモデルでは, ユーザーがモーラ単位でアクセントラベルを調整する必要があり, フィードバックを与える難易度が高いという問題がある.

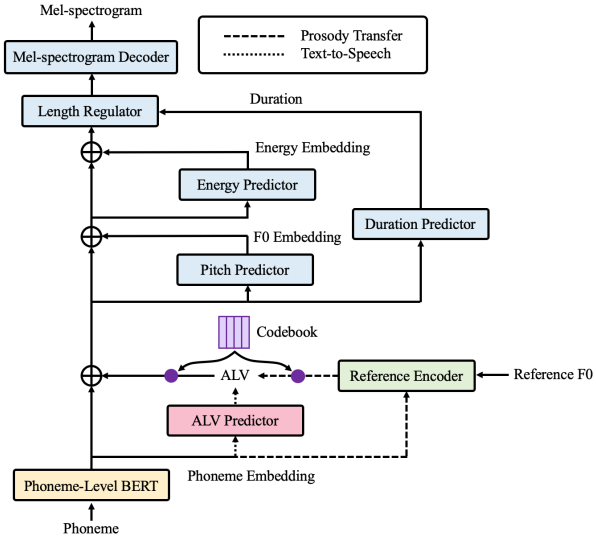

Fig. 1　提案手法の TTS モデルに関する概要図.

3　提案手法

提案モデルは, Yufune らの音声から自動で ALV を抽出することができる TTS モデル [2] と, Fujii らのアクセント誤り訂正可能な TTS モデル [3] を統合したモデルと捉えることができる. Fig. 1 に提案モデルの概略図を示す. モデルのベースには FastSpeech2 [4] を用いた. 提案モデルは FastSpeech2 に, 1) 音素エンコーダーとして Phoneme-Level BERT(PL-BERT) [5] を使用, 2) 音素列のみから ALV 列を予測する ALV Predictor を導入, 3) 参照音声から ALV を抽出する Reference Encoder を導入するといった変更を加えたモデルである. ALV Predictor により音素列から ALV 列を予測することができるが, ユーザーが参照音声を入力することで望ましい韻律の合成音声を得ることもできる (Prosody Transfer).

3.1　Reference Encoder

Reference Encoder は, 参照音声を入力として受け取り基本周波数 (F0) の埋め込み表現を出力する Reference Pitch Encoder と, F0 埋め込み列と音素埋め込み列を入力として受け取り ALV 列を出力する VQ-VAE Encoder からなる.

VQ-VAE Encoder は F0 埋め込みと音素埋め込みの和を連続な潜在表現に埋め込んだ後, それを特定のクラス数に量子化する. 本研究では, 量子化する際のクラス数は 4 とする. また, 量子化された F0 埋め込みと音素埋め込みの潜在表現を ALV と呼ぶ.

学習時は, 単に教師データであるターゲット音声を参照音声として使用する. Reference Encoder は FastSpeech2 の損失関数によって, その他のモジュールと共同で学習される.

---

*Investigation of dialect speech synthesis using a TTS model that can predict and control Accent Latent Variable, by YAMAUCHI Kazuki, SAITO Yuki, SARUWATARI Hiroshi (The University of Tokyo).

3.2 ALV Predictor

ALV Predictor は, 音素列を入力として受け取り ALV 列を返すモデルである. テキストのみから ALV を予測するには, 十分な語彙を含むデータで学習する必要がある. しかし, 方言音声コーパスからは十分な語彙を獲得できないため, テキストと ALV のペアデータを十分用意することは難しい. そこで, 音素エンコーダーとして, テキストのみで事前学習することで TTS モデルの自然性と韻律を向上させる文脈埋め込みを生成する音素単位の言語モデルである, PL-BERT を用いた. PL-BERT は 2 段階で学習される.

第 1 段階では, masked phoneme token prediction task および phoneme-to-grapheme prediction task によって事前学習される. ソフトマックス関数と線形射影を使用し, 最終層の隠れ状態からマスクされた入力音素および各音素に対応する単語を予測する.

第 2 段階では, PL-BERT は他のモジュールと共同で学習される. ALV Predictor は, PL-BERT の最終層の隠れ状態から, 線形射影を使用して ALV を予測するモデルである. これは, Reference Encoder の出力 ALV と ALV Predictor の予測 ALV とのクロスエントロピー誤差を損失として学習される.

4 評価実験

4.1 TTS モデルの実験条件

本実験では, 単一女性話者により構成される JSUT コーパス [6] と, 大阪方言と熊本方言を含む多方言音声コーパスである JMD コーパス [7] を用いた. また, F0 分析には WORLD [8] を用い, 音素アライメント情報は Julius [9] で取得した. ボコーダには HiFi-GAN [10] を使用した. また, PL-BERT の事前学習には日本語 Wikipedia コーパス [1]を用いた.

本実験で比較するモデルは以下の 3 つである. 各モデルはまず JSUT コーパス (約 7700 発話) により学習した後, JMD コーパスの大阪方言サブセット (1300 発話) を用いてファインチューニングした.

- **FS2 w/o Acc**: FastSpeech2 にアクセント情報を与えず学習した手法
- **FS2 w/ AP**: FastSpeech2 を提案手法の枠組みで学習し, ALV Predictor によってテキストから予測された ALV を用いる手法
- **FS2 w/ PT**: FastSpeech2 を提案手法の枠組みで学習し, Reference Encoder によって参照音声から抽出された ALV を用いる手法

クラウドソーシングを用いて, 音声の自然性および大阪方言らしさに関する MOS テストを実施した. 各手法の合成音声をランダムに提示し, 音声の自然性を 5 段階, 大阪方言らしさを 3 段階で評価させた. この評価での受聴者数は 40 人で, 1 人の評価回数は 24 である. FS2 w/ PT の参照音声には JMD コーパスのテストセットを用いた. また, 比較のため JMD コーパスの生音声も用いた. 結果を Table 1 に示す.

FS2 w/ PT の自然性および大阪方言らしさに関する MOS スコアは, FS2 w/o Acc よりも有意に高くなっている. これは, 望ましい ALV を入力すること

[1] https://dumps.wikimedia.org/

Table 1 合成音声の自然性および大阪方言らしさに関する MOS スコア (±95% 信頼区間)

手法	自然性 MOS	方言性 MOS
JMD	4.57 ± 0.071	2.75 ± 0.065
FS2 w/o Acc	2.95 ± 0.117	2.08 ± 0.081
FS2 w/ AP	2.71 ± 0.102	1.75 ± 0.079
FS2 w/ PT	3.19 ± 0.118	2.28 ± 0.077

で, 合成音声の自然性および方言らしさを向上させることができることを示す. 一方, FS2 w/ AP の MOS スコアは FS2 w/o Acc よりも有意に低くなっている. これは, 望ましくない ALV が入力されることで, 合成音声の自然性および方言らしさが著しく下がってしまうということを示す. この原因としては, 1) 学習時に ALV Predictor の出力が合成音声にはほぼ反映されないため, 2) データ数が少なく ALV Predictor の精度が十分でないためといった要因が考えられる.

5 おわりに

本研究では, アクセント潜在変数 (ALV) の予測と制御が可能な TTS モデルを提案し, 提案モデルによる方言音声合成を検討した. 実験的評価により, 望ましい韻律の参照音声を入力し適切な ALV を入力することで, 合成音声の自然性および方言らしさを向上させることができることがわかった. 今後は, パブリックユーザーによるフィードバックを用いた ALV Predictor の継続学習について検討する予定である.

謝辞: 本研究は公益財団法人 立石科学技術振興財団 2023 年度研究助成 (S) による支援を受けたものです.

参考文献

[1] A. van den Oord et al., "Neural discrete representation learning," in *Proc. NIPS*, vol. 31, Long Beach, California, USA, Dec. 2017, pp. 6309—-6318.

[2] K. Yufune et al., "Accent modeling of low-resourced dialect in pitch accent language using variational autoencoder," in *Proc. SSW*, Budapest, Hungary, Aug. 2021, pp. 189–194.

[3] K. Fujii et al., "Adaptive end-to-end text-to-speech synthesis based on error correction feedback from humans," in *Proc. APSIPA ASC*, Chiang Mai, Thailand, Nov. 2022, pp. 1702–1707.

[4] Y. Ren et al., "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, Vienna, Austria, May 2021.

[5] Y. A. Li et al., "Phoneme-level BERT for enhanced prosody of text-to-speech with grapheme predictions," vol. abs/2301.08810, 2023. [Online]. Available: https://arxiv.org/abs/2301.08810

[6] S. Takamichi et al., "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, Sep. 2020.

[7] S. Takamichi, H. Saruwatari, "JMD: Japanese multi-dialect corpus," 2021. [Online]. Available: https://sites.google.com/site/shinnosuketakamichi/research-topics/jmd_corpus?authuser=0

[8] M. Morise et al., "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, Jul. 2016.

[9] A. Lee et al., "Julius — an open source real-time large vocabulary recognition engine," in *Proc. EUROSPEECH*, Aalborg, Denmark, Sep. 2001, pp. 1691–1694.

[10] J. Kong et al., "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, vol. 33, Virtual Conference, Dec. 2020, pp. 17 022–17 033.