E-010

拍・アクセント辞書を明示条件とした HiFi-GAN 日本語音声合成

Japanese Speech Synthesis Using HiFi GAN Explicitly Conditioned on a Mora-and-Accent Dictionary

高林 竜久斗 † 李 嘉誠 † 能登 正人 †
Rikuto Takabayashi Jiacheng Li Masato Noto

1. はじめに

近年，音声合成技術は対話型 AI アシスタント，自動字幕生成，ナビゲーションシステム，教育・福祉支援ツールなど，様々な場面で社会実装が進んでいる．特に深層学習技術の急速な発展を背景に，従来の統計的手法に比べ，音質・表現力ともに格段に向上した音声波形生成モデルが次々と登場している．一方で日本語音声合成では，「拍（mora）」や「高低アクセント」といった独特の韻律構造を持つため，リズムやイントネーションの再現が極めて難しく，その再現度が合成音声の自然性・明瞭度・知覚品質を大きく左右する．従来の日本語 TTS (Text-to-Speech) システムでは，これらの音韻情報を入力特徴量から暗黙的に学習させることが主流だったが，言語特有の情報を十分に活かしきれず，イントネーション誤りや不自然な発話リズムが残るなど，実用面で課題が多かった．

本研究では，高速かつ高音質な波形生成を可能とする HiFi-GAN[1] に着目し，日本語音声合成における「拍」や「アクセント」辞書情報を明示的に特徴量としてモデルに与える新たなアプローチを提案する．この手法の有効性を，客観的な音響指標や系列別誤差分析など多角的な観点から詳細に検証することを目的とする．

2. 先行研究

深層学習の導入以降，音声合成分野は急速な発展を遂げている．WaveNet[2] は，ボコーダ方式を超える自然性と柔軟性を実現し，音声合成の新たな標準となった．また Tacotron 系 [3] は，テキストから音響特徴量へのエンドツーエンド生成を実現し，合成音声の流暢さや表現力を大きく向上させた．しかし，WaveNet は逐次生成の構造により推論速度が遅く，Tacotron も音響特徴量生成やリアルタイム性，複雑な韻律制御に課題があった．こうした背景から，近年は GAN に基づく HiFi-GAN など，高速かつ高音質な波形生成を可能にする手法が台頭し，実用 TTS システムにも数多く導入されている．

一方，日本語 TTS におけるリズム・アクセント再現の重要性は，早くから指摘されてきた．英語と比較して，日本語は「拍」単位で話されるリズムや，単語ごとに割り当てられる「高低アクセント」が語義区別や聞き取りやすさに直結する特徴を持つ．従来の深層学習ベース TTS モデルでは，こうした日本語特有の韻律情報を入力系列や音響特徴量から暗黙的に学習させるケースが多かった．しかし，十分な再現が難しく，リズムの乱れやアクセント誤りが合成音声の知覚品質を損なう要因となっていた．近年は，拍やアクセント情報を明示的に辞書から抽出し，モデルの入力特徴量として統合する手法も提案されているが，高速波形生成モデルとの連携やその効果検証はまだ十分でない．

† 神奈川大学，Kanagawa University

3. 提案手法

本研究では，HiFi-GAN の Generator に対し，日本語 TTS に特有な韻律情報である拍・アクセント辞書情報を 16 次元特徴ベクトルとして並列入力する拡張構成を提案する．これにより，モデルは音響特徴量と同時にフレーム単位で拍やアクセントの変動を直接学習可能となり，より自然な日本語音声合成を実現できる．

図 1 に提案システム全体構成の概要を示す．

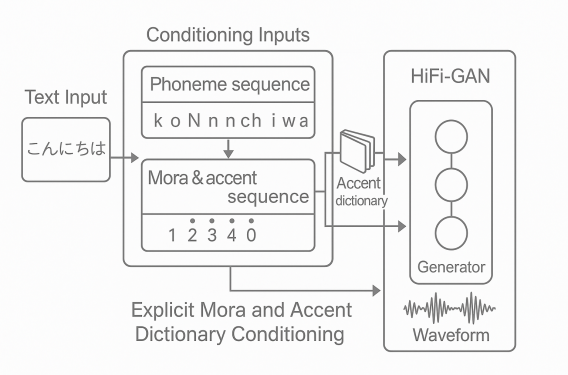

図 1: 提案システムの構成概要

4. 実験

4.1 データセット・前処理

実験には JSUT コーパスの単一話者全発話データを利用した．各 wav ファイルに対して PyOpenJTalk-plus を適用し，音素・拍・アクセント系列を高精度かつ自動で抽出した．抽出後，各系列と音声データの整合性や欠損を独自スクリプトにより厳格に検査し，不一致や欠損の見られるファイルはすべて除外した．音声データは 24kHz/16bit で正規化し，音素・拍・アクセント配列を

NumPy 形式で時系列フレームと完全対応させて保存した．この工程により，各特徴量がフレーム単位で一対一に対応し，モデル入力時の一貫性が保証された．

4.2 学習・モデル構築

構築したデータセットを，学習（2000 発話），検証（500 発話），テスト（500 発話）の 3 分割で利用した．音響モデルは HiFi-GAN v1 構造を基本とし，提案した辞書ベース特徴量を並列入力できるよう Generator を拡張した．学習パラメータは表 1 に示すとおり，エポック数，バッチサイズ，オプティマイザ（AdamW）などを日本語音声に最適化した．学習時には Early Stopping やデータシャッフル，重み初期化，損失カーブの逐次モニタリングなど，学習安定化や過学習抑制の工夫も実施した．

表 1: 主な学習パラメータ

パラメータ	設定値
サンプリングレート	24,000 Hz
バッチサイズ	16
セグメント長	8192
エポック数	500
Optimizer	AdamW
Mel 次元	80

4.3 評価方法

性能評価には，Mel スペクトログラム L1 誤差（MelL1），短時間明瞭度指標（STOI），メルケプストラム歪み（MCD），対数スペクトル距離（LSD）といった，音声合成分野で広く用いられている標準的な客観指標を採用した．MelL1 は合成音声と教師データ間のメルスペクトログラム差分を L1 ノルムで定量化するもので，主にスペクトル再現性を評価できる．STOI は音声の明瞭度・可聴性を客観的に示す指標であり，TTS の実用品質を判定する重要な基準である．MCD と LSD は音響信号のスペクトル的な歪みを細かく測定でき，発話の自然性や類似度の観点から多面的な性能比較が可能となる．

本研究では，学習・検証・テストそれぞれのデータセットについて各指標を定期的に算出し，エポックごとの変化を詳細に記録した．さらに，単なる平均値比較だけでなく，拍やアクセント型といった系列単位で誤差分布の解析も行い，拍・アクセント情報を明示的に条件付けした場合の効果を多角的に検証した．

5. 結果・考察

各種指標のエポックごとの推移を表 2 に示す．学習初期（100 エポック）では MelL1，MCD，LSD が高かったが，エポックが進むごとに着実に減少し，500 エポック時点では MelL1=0.289，MCD=5.08，LSD=1.91，STOI=0.924 と，音響的な再現性・明瞭度の両面で良好な値となった．

これらの結果から，提案手法が従来の暗黙的学習や単純前処理のモデルに比べ，合成音声の知覚明瞭度や韻律の自然さを向上させていることが明らかになった．特に STOI が 0.92 を超えたことは，聞き取りやすさの大幅な改善を示し，拍・アクセント系列を明示的に導入した意義が示唆される．また，拍・アクセント型ごとの誤差集中が緩和されたことから，日本語 TTS で問題となるリズム・イントネーションの不自然さに対して，本研究のアプローチが有効であると考えられる．

さらに，本研究の目的である「日本語に特化した拍・アクセント辞書情報の明示的活用が音声品質に与える影響の検証」に関して，複数の客観指標で改善効果が確認されたことから，目的を十分に達成できたと判断できる．ただし，今回は単一話者・限定的な条件下での検証に留まっているため，より多様な話者や発話条件，主観評価の導入など，今後の課題も残る．

表 2: エポックごとの評価指標

Epoch	MelL1	STOI	MCD	LSD
100	0.452	0.893	5.72	2.10
300	0.317	0.917	5.28	1.97
500	0.289	0.924	5.08	1.91

6. おわりに

本研究では，日本語 TTS における拍・アクセント辞書情報を明示的特徴量として HiFi-GAN に入力し，客観的な音響指標を用いてその有効性を詳細に検証した．その結果，MelL1，MCD，LSD の誤差低減や STOI の明瞭度向上など，音響的品質と知覚品質の双方で顕著な改善が認められ，本手法の有効性が示された．一方，主観評価（MOS）の導入，多話者・感情音声拡張，より多様な発話条件での適用検証，さらなる韻律特徴量の統合など，今後の発展的課題も多い．今後は，これらの課題に取り組むことで，日本語音声合成技術のさらなる高度化と実用化に貢献する．

参考文献

[1] Kong, J., Kim, J. and Bae, J.: HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis, *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 17022–17033 (2020).

[2] van den Oord, A. et al.: WaveNet: A Generative Model for Raw Audio, *Proceedings of 9th ISCA Speech Synthesis Workshop*, pp. 125–129 (2016).

[3] Wang, Y. et al.: Tacotron: Towards End-to-End Speech Synthesis, *Proceedings of 18th Annual Conference of the International Speech Communication Association*, pp. 4006–4010 (2017).