HiFi-GAN ボコーダにおける LPCNet 特徴量の検討 *

◎ 松原圭亮[1,2]，岡本拓磨[2]，高島遼一[1]，滝口哲也[1]，戸田智基[3,2]，河井恒[2]
[1] 神戸大学，[2] 情報通信研究機構，[3] 名古屋大学

1 はじめに

テキスト音声合成 (Text-to-Speech: TTS) や声質変換は音声コミュニケーションの重要な技術である。近年では，深層学習を用いた品質の向上がめざましく，自然音声に近い高品質な音声を生成できるようになっている [1]。これらの発展の大きな転換点として WaveNet ボコーダ [2] をはじめとするニューラルボコーダの登場がある。ニューラルボコーダはメルスペクトログラムなどの音響特徴量から音声を復元するボコーダに深層学習を適用したもので，従来のソースフィルタボコーダ [3] による品質を大きく上回り，ニューラル音声合成技術の発展に大きく貢献している。

WaveNet ボコーダには合成速度が遅いという問題があったが，今日では様々なモデルが提案され，高品質な音声をリアルタイムに合成できるようになってきている。またニューラルボコーダは，一般的に学習に用いていない未知話者の音声を合成すると品質が劣化するが，大量の複数話者の音声を学習に用いることでその課題を解決する試みもなされている [4]。これらのアプローチは複数話者 TTS や多対多の声質変換モデルと組み合わせることでそれらの音声の品質を向上させることができる。近年では，複数話者音声合成を実現しながら CPU でリアルタイムに動作可能なほど軽量なモデルが提案されている [5, 6]。

本稿では，HiFi-GAN [5] や MWDLP (Multiband WaveRNN with data-driven linear prediction) [6] の高速合成が可能な複数話者ニューラルボコーダと，文献 [7] にて複数話者合成が報告されている LPCNet [8] について，合成音声の品質および合成速度について比較，考察する。また，LPCNet で用いている音響特徴量は TTS 音響モデルの学習に用いた際に劣化が少なく頑健性が高いことが報告されており [9]，この LPCNet 特徴量を HiFi-GAN に適用した場合の検討も行う。

2 高速複数話者ニューラルボコーダ

本稿では，HiFi-GAN, MWDLP，LPCNet, の 3 つのニューラルボコーダを採用した。HiFi-GAN, MWDLP は高速かつ複数話者音声合成が可能なニューラルボコーダとして提案されており，また LPCNet も同様に複数話者音声合成が可能と報告されている。なお，自己回帰型である MWDLP および LPCNet に関しては低遅延リアルタイム合成も容易であるが，本稿ではバッチ処理を想定したリアルタイム合成に着目する。

2.1 LPCNet

LPCNet は WaveRNN [10] をベースとしたニューラルボコーダであり，線形予測分析 (Linear Predinction Coding: LPC) で推定した音声信号の残差信号をニューラルネットで推定する。LPCNet は，入力された音響特徴量から特徴量抽出を行う Frame Rate Network と，LPC で推定された音声信号から残差信号を推定する Sample Rate Network の 2 つのブロックで構成されている。ここで，入力特徴量はバーク尺度のケプストラムとピッチ周期，ピッチ相関である。LPCNet では，モデルパラメータ数の削減と Sample rate network に用いられる Gated Recurrent Unit (GRU) に対して，Sparse coding を適用することで高速化を行っている.

また文献 [11] にてシングルコア CPU によるリアルタイム合成が，文献 [7] にて複数話者音声合成が可能と報告されている。

2.2 MWDLP

MWDLP も LPCNet と同様に WaveRNN [10] をベースとして, 品質及び合成速度を改善したニューラルボコーダである。WaveRNN は音声波形予測を 16 bit 信号の離散型分類問題として捉え，下位 8 bit を推定する GRU と上位 8 bit を推定する fine GRU の 2 つの GRU のみで構成されている。このように WaveRNN はモデル構造をコンパクトにすることで高速な合成を実現しているが，音声の品質については WaveNet などの巨大なモデルを用いたニューラルボコーダと比較して劣っていた。MWDLP では WaveRNN の構造をベースとして，μ-law 圧縮を用いた推定ビット数の削減，マルチバンド処理の導入による高速化，線形予測分析の導入などによって，音声品質およびリアルタイム性の向上を達成しており，未知話者に対する高品質な音声合成とシングルコア CPU におけるリアルタイム合成を実現している [6]。

---

*Investigation of LPCNet features in HiFi-GAN vocoder by MATSUBARA, Keisuke[1,2], OKAMOTO, Takuma[2], TAKASHIMA, Ryoichi[1], TAKIGUCHI, Tetsuya[1], TODA, Tomoki[3,2], and KAWAI, Hisashi[2] ([1]Kobe Univ, [2]NICT, [3]Nagoya Univ)

2.3 HiFi-GAN

HiFi-GAN は敵対的生成ネットワークをベースとするニューラルボコーダであり，転置畳み込みを用いて入力特徴量を音声信号に変換する生成器と複数のサンプリング周波数と受容野を持つ 2 種類の識別器から構成される。

生成器は通常の畳み込み層と転置畳み込み層 (Transposed convolution) から構成されており，入力された音響特徴量を転置畳み込みを用いてアップサンプリングしながら音声波形に変換する。また異なる受容野を持つ複数の畳み込み層の出力を足し合わせる処理を行うことで，音声信号の様々な周期成分を捉えやすくする処理が施されている。

識別器は複数のサンプリング周波数において合成音声の真偽を識別する Multi-Scale Discriminator (MSD) と，音声信号を様々な間隔でサンプリングしてそれらの信号から真偽を識別する Multi-Period Discriminator (MPD) から構成される。MSD は出力音声にダウンサンプリングを施し，数種類の異なるサンプリング周波数の信号に対して別々の識別器で識別する。MPD では長さ T の音声信号に対して間隔 d でサンプリングを行い，$(T/d) \times d$ の 2 次元信号に変形した後に識別器に入力する。そして間隔 d を複数個設定し，各々において別々の識別器を用いて学習を行う。これらの処理により，音声信号に含まれている様々な周期成分を効率的に捉えることが可能となっている。結果として，小さいサイズの生成器でも高品質な合成が可能となっており，CPU でのリアルタイム合成を実現している [5]。

3 LPCNet 特徴量を用いた HiFi-GAN

本稿では HiFi-GAN に対して前述の LPCNet 特徴量の利用を検討する。従来では，HiFi-GAN は入力特徴量に 80 次元のメルスペクトログラムを使用している。LPCNet 特徴量はサンプリング周波数 24 kHz の場合，30 次元のバークケプストラムとピッチ周期，ピッチ相関の計 32 次元の特徴量である [9]。LPCNet 特徴量は低次元の特徴量であることから音響モデルで推定した際の劣化が小さく，また声道特徴と声帯特徴が分離された特徴量であることから制御が容易であるなどのメリットがある。文献 [9] では，TTS 音響モデルと組み合わせた際に品質の劣化が少なく，LPCNet 特徴量が音響モデルによる劣化に対して頑健であることが報告されている。また MWDLP に対しても LPCNet 特徴量での事前実験を行ったが，正常に合成が行えなかった。これは MWDLP がメルスペクトログラム用にモデルパラメータ等が調整されている可能性が高く，詳細な調査は今後の課題とする。

4 実験

4.1 実験条件

HiFi-GAN，MWDLP，LPCNet における合成音声の品質を評価するため，サンプリング周波数 24 kHz の音声を用いた分析合成およびテキスト音声合成での主観評価を行う。比較対象には LPCNet を用いた。データセットは JSUT コーパス [12] より，日本人女性話者 1 名による 7,697 文の音声と，JVS コーパス [12] より，100 名の日本人話者による各話者 130 文の音声を使用した。学習には JSUT コーパスを用いた単一話者学習と JVS コーパスを用いた複数話者学習の 2 種類を行った。単一話者学習では 7,497 文を学習に用いて，残りの 200 文の内 100 文ずつを評価および検証に用いた。複数話者学習では 95 名の 12,350 文を学習に用い，4 名の 120 文の音声を検証に，男性 1 名 (jvs001) および JSUT 話者の 60 文の音声を評価に用いた。またニューラル音響モデルの学習には JSUT コーパスの内，人手によりラベリングされた基づく HTS 形式のコンテキストラベルが利用可能な 4,800 文を使用した。1 テストセットと検証セットにはそれぞれ 100 文ずつを使用した。

TTS では，フルコンテキストラベルを入力とする FastSpeech 型の音響モデルを用いた [13, 14]。入力特徴量は 38 次元の音素と HTS 形式のフルコンテキストラベルの A と F から 9 次元のアクセントラベルを抽出した 47 次元のベクトルとした。

HiFi-GAN は [5] による実装の内、モデルサイズの最も大きい V1 と 2 番目に大きい V2 の 2 つを使用した。入力特徴量にはメルスペクトログラム 80 次元と，LPCNet で用いられているバークケプストラム 30 次元とピッチ周期，ピッチ相関の計 32 次元特徴量の 2 種類で実験を行った。メルスペクトログラムの計算時は，計算では窓長を 42.7 ms，フレームシフトを 10 ms とし，周波数帯域を 80〜7.6 kHz に制限した。バークケプストラムの計算に際しては，窓長を 20 ms，フレームシフトを 10 ms としてスペクトル分析を行い，バーク尺度によるフィルタバンクを適用した後に離散コサイン変換を行った。ピッチの計算にはオープンループの相互相関関数をベースとする手法を用いた。従来の HiFi-GAN は 256 倍のアップサンプリングを導入しているが [5]，本検討ではフレームシフトが 10 ms のため，240 倍のアップサンプリングなる。したがって，アップサンプリング数を $[5, 4, 3, 4]$ とし，転置畳み込みのカーネルサイズを $[11, 8, 7, 8]$ とした。また事前実験の結果、TTS 音響モデルから出力された特徴量を入力した際の品質の劣化が顕著であったため，音響モデルから出力される訓練データの音響特徴量を用いてファインチューニ

ングを行った。ファインチューニングでは，100 万ステップ学習を行ったモデルに対して 20 万ステップの追加学習を行った。

MWDLP は [6] と同様のモデルを使用した。入力特徴量にはメルスペクトログラム 80 次元を使用し，計算では窓長を 42.7 ms，フレームシフトを 10 ms とした。マルチバンド処理の帯域分割数は 6，LPC 分析の次元数を 8 とした。また TTS についても HiFi-GAN と同様に音響モデルからの特徴量を用いてファインチューニングを行ったが品質の改善が見られなかったため本実験の条件から除外した。LPCNet は [8] と同様のモデルを使用した。入力特徴量は HiFi-GAN で使用したものと同じ 32 次元の LPCNet 特徴量を用いた。

4.2 実験結果

4.2.1 合成速度の比較

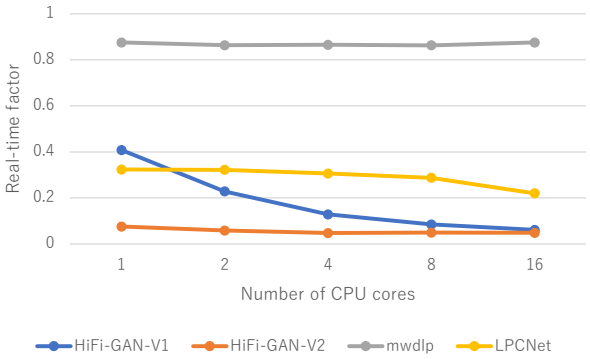

Fig. 1 Result of real-time factors for inference using Intel Xeon 6152.

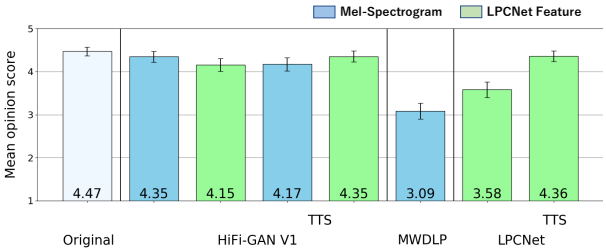

Fig. 2 Result of MOS test with the single-speaker model.

Fig. 1 に各モデルにおける CPU コア数による合成速度の変化を示す。HiFi-GAN は入力特徴量が 2 条件あるが，どちらの条件でも合成速度に差は出なかったためメルスペクトログラムを用いたモデルでの結果を示している。図より，いずれのモデルもシングルコア CPU の RTF は 1 を下回り，リアルタイム合成が可能であると確認できた。また前段の FastSpeech 型音響モデルの合成速度 (RTF=0.05 [9]) を合わせてもリアルタイム性が維持されていることも確認できた。HiFi-GAN V1 を除くモデルは CPU コア数の増加における速度の改善は微量であった。MWDLP および LPCNet については，GRU を用いた自己回帰モデルであることからコア数が増加しても並列処理が行われなかったためだと考えられる。HiFi-GAN V2 についてはモデルサイズが軽量のため，シングルコアの時点で RTF=0.08 となっておりコア数を 16 に増加させた場合でも RTF=0.05 と変化は微少となった。HiFi-GAN V1 はコア数の変化による速度の改善が顕著に見られ，シングルコアで RTF=0.40 から，16 コアで RTF=0.06 まで改善された。従来の MWDLP はシングルコアで RTF が 0.6 程度であるが，Fig. 1 では 0.87 となった。これを受けて別の CPU (Intel Core i5-9400) で計測した結果、RTF=0.64 となり，CPU の種類などの要因に大きく依存する可能性があることが分かった。これらの詳しい調査は今後の課題とする。

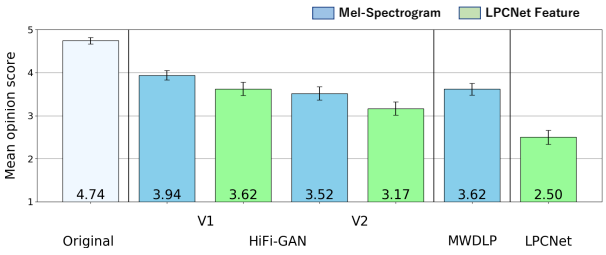

Fig. 3 Result of MOS test with multi-speaker model using speech of unseen female speaker.

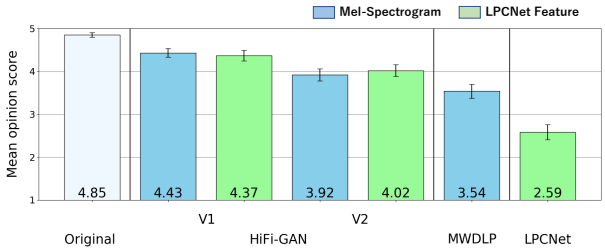

Fig. 4 Result of MOS test with multi-speaker model using speech of unseen male speaker.

4.2.2 主観評価実験の結果

主観評価として，聴取実験による平均オピニオン評点テストを行った。実験参加者は健常な聴覚である 10 人の成人日本語母語話者で，テストセット 18 文に対して 22 条件の計 396 文をヘッドホン聴取により評価した。また FastSpeech によるテキスト音声合成の実験は，HiFi-GAN および LPCNet に対して行った。

Fig. 2 に単一話者学習を行った場合の結果を示す。図より，分析合成では HiFi-GAN V1 が最も高いスコアを示した。また入力する音響特徴における品質の違いは見られなかった。MWDLP や LPCNet は，原音と比較して掠れ成分が多くなる傾向が見られ，それに応じてスコアが低くなったと考えられる。また

MWDLP はマルチバンド処理をする関係で帯域の境界付近の周波数成分をもつ雑音が微量ではあるが発生しており，それらもスコアが低くなった原因と考えられる。テキスト音声合成では，HiFi-GAN は分析合成と比較して有意差がなく，LPCNet は分析合成よりも高いスコアを示した。HiFi-GAN については 4.1 で述べた通り，TTS 音響モデルから出力される特徴量でファインチューニングを行っている。ファインチューニングを行わなかった場合は品質が著しく劣化することが確認出来たため，このような劣化した音響特徴に対する頑健性を高めることは今後の課題の 1 つと言える。LPCNet は分析合成よりも高いスコアとなったが，これは TTS 音響モデルから出力された特徴量が高周波成分が平滑化される傾向にあることから，上述した LPCNet の合成音に見られた掠れ成分の増加が知覚上抑えられたためと考えられる。

Fig. 3 および Fig. 4 に複数話者学習を行ったモデルによる分析合成の結果を示す。単一話者学習の結果と比較して，全体的にスコアが低くなる傾向が見られたが，HiFi-GAN と MWDLP は LPCNet に対して有意なスコアを達成した。HiFi-GAN はいずれの条件でも V1 の方が V2 より高いスコアを示した。Fig. 3 の女性話者での実験では，入力特徴量にメルスペクログラムを用いた場合の方が LPCNet 特徴量を用いた場合と比較して有意なスコアを示した。これは LPCNet 特徴量の方が 32 次元と次元が小さく情報量が少ないため，未知話者の音声に対する頑健性が低下したためだと考えられる。HiFi-GAN と MWDLP を比較すると，Fig. 3 の女性話者での実験で HiFi-GAN V2 と MWDLP が同等のスコアとなり，それ以外の条件ではすべて HiFi-GAN の方が高いスコアを示した。

5 おわりに

HiFi-GAN，MWDLP，LPCNet ボコーダについて合成速度および音声の品質について評価を行った。合成速度についてはいずれもリアルタイム合成が可能であると確認ができ，品質については HiFi-GAN がほとんどの場合で高い品質を示した。また HiFi-GAN での LPCNet 特徴量を用いた実験では，一部の条件を除いてメルスペクトログラムの場合と同等の品質を示した。今後は劣化が確認された条件について原因の分析や品質の改善について検討する。

参考文献

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

[2] A. Tamamori *et al.*, “Speaker-dependent WaveNet vocoder,” in *Proc. Interspeech*, Aug. 2017, pp. 1118–1122.

[3] M. Morise *et al.*, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE trans, Inf. Syst.*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[4] J. L.-Trueba *et al.*, “Towards achieving robust universal neural vocoding,” in *Proc. Interspeech*, July 2019, pp. 181–185.

[5] J. Kong, *et al.*, “HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis,” in *Proc. NeurIPS*, Dec. 2020, pp. 17022–17033.

[6] P. L. Tobing, and T. Toda, “High-fidelity and low-latency universal neural vocoder based on multiband WaveRNN with data-driven linear prediction for discrete waveform modeling,” in *Proc. Interspeech*, Aug. 2021.

[7] K. Matsubara *et al.*, “High-intelligibility speech synthesis for dysarthric speakers with LPCNet-based TTS and CycleVAE-based VC,” *Proc. ICASSP*, June 2021, pp. 7058–7062.

[8] J. Valin, and J Skoglund, “LPCNet: improving neural speech synthesis through linear prediction,” *Proc. ICASSP*, May 2019, pp. 5891–5895.

[9] K. Matsubara *et al.*, “Full-band LPCNet: A real-time neural vocoder for 48 kHz audio with a CPU,” in *IEEE Access*, Vol. 9, pp. 94923–94933, 2021.

[10] N. Kalchbrenner *et al.*, “Efficient neutal audio synthesis,” in *Proc. ICML*, July 2018, pp. 2415–2424.

[11] K. Matsubara *et al.*, “Investigation of training data size for real-time neural vocoders on CPUs” *Acoust. Sci. Tech.*, vol. 42, no. 1, pp. 65–68, Jan. 2021.

[12] S. Takamichi *et al.*, “JSUT and JVS: free Japanese voice corpora for accelerating speech synthesis research,” *Acoust. Sci. Tech.*, vol. 41, no. 5, pp. 761–768, Sept. 2020.

[13] Y. Ren *et al.*, “FastSpeech: fast, robust and controllable Text to Speech” in *Proc. NeuIPS*, Dec. 2019, pp. 3165–3174.

[14] T. Hayashi, *et al.*, “ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit,” in *Proc. ICASSP*, May 2020, pp. 7654–7658.