Period-HiFi-GAN: 基本周波数を制御可能な高速ニューラルボコーダ *

◎ 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のひとつであり，近年では深層学習を用いた品質の向上によって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。特に盛んに研究されている手法として，パラレル生成モデルと敵対的学習を用いた手法がある。これらの手法ではパラレル生成モデルによって複数の音声サンプルを同時に出力することで高速合成を行いながら，敵対的学習 [4] を導入することで自然性の高い音声の合成を可能にしている [5, 6, 7]。

しかし，これらのニューラルボコーダはデータ駆動型のため，基本周波数 (F0) に対する忠実度および制御性能において従来のソースフィルタボコーダに劣っているという課題点があった。この課題を解決するために，ソースフィルタボコーダと同様に F0 に対応した励起信号を入力する手法が提案されている [8, 9]。また PeriodNet では，歌声合成等のピッチの変動が大きい音声を合成する場合でも同様の手法が有効に働くことが報告されている [10]。しかし，これらの手法はリアルタイム合成のためにはハイエンドな GPU が必要であることや，PeriodNet は通常音声の合成が著しく劣化するという問題があった [11]。

本研究では，高速かつ高品質なニューラルボコーダとして提案されている HiFi-GAN [7] に対して励起信号を入力するネットワークを導入した Period-HiFi-GAN を提案する。具体的には，HiFi-GAN の生成器が入力音響特徴量を音声信号のサンプリング周波数まで段階的にアップサンプリングしていくのに加え，提案法では F0 に対応した励起信号を段階的にダウンサンプリングしていく層を新たに追加する。実験では未知話者音声の分析合成，および F0 をスケーリングした場合についての評価実験を行い提案法の性能を評価する。

2 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，音声信号の特徴を効率的に捉えるための 2 つの識別器を提案している。

生成器は通常の畳み込み層および転置畳み込み層 (Transposed convolution) から構成され，入力の音響特徴量を転置畳み込みを用いて段階的にアップサンプリングしながら音声波形に変換する。Parallel WaveGAN 等の従来のニューラルボコーダが数十段の畳み込み層を用いて音声波形を生成するのに対して，HiFi-GAN では数段程度の層数で生成器を設計することで CPU のみを用いたリアルタイム生成を実現している。

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) [6] 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。その上で間隔 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，畳み込み層数の少ない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 [7, 12, 13]。

また，HiFi-GAN は，メルスペクトログラムではなく LPCNet [14] 特徴量のようなソースフィルタ型ボコーダの特徴量でも頑健に動作することが示されている [13]。

---

* Period-HiFi-GAN: Fast and fundamental frequency controllable neural 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)

3　提案法：Period-HiFi-GAN

本研究では，HiFi-GAN の生成器に対して励起信号を入力する層を新たに導入する。Fig. 1 に提案する Period-HiFi-GAN の生成器の概要を示す。ここでは，入力特徴量にメルケプストラム (Melcep)，非周期性指標 (Bap) および声門閉鎖点 (GCI) を用いる。メルケプストラムと非周期性指標はアップリング層に入力され，HiFi-GAN と同様に転置畳み込みを用いて音声波形へと変換される。ここでアップリング層は，転置畳み込み層および HiFi-GAN で提案されているマルチ受容野混合層 (MRF) を連結した T.Conv ブロック数段で構成される [7]。声門閉鎖点は正弦波生成器に入力され，教師音声信号のピッチに対応した正弦波に変換された後にダウンサンプリング層に入力される。ダウンサンプリング層は数段の畳み込み層で構成され，入力の正弦波を段階的にダウンサンプリングしながら，各畳み込み層の出力がアップサンプリング層の各ブロックの出力に加えられる。この手法を導入することで，出力音声信号が励起信号との整合性を保つよう学習されることが期待される。一般的にピッチ特徴量には基本周波数が用いられることが多いが，声門閉鎖点は音声信号の位相情報も含まれているため，励起信号を生成する際に教師音声信号との位相を合わせることができるというメリットがある [10]。

ここでは，声門閉鎖点の秒数を並べた系列にサンプリング周波数 F_s を掛けたものを $\mathbf{g} = [g_1, \cdots, g_n, \cdots, g_N]$ とする。各 g_n は声門が閉鎖するタイミングが何フレーム目かを示している。そして，$\mathbf{vuv} = [v_1, \cdots, v_T]$ を各時刻における有声/無声シンボル系列とする。このとき，励起信号 $\mathbf{e} = [e_1, \cdots, e_T]$ は以下のように生成される。

$$k = \arg \min_{\{n:g_n<t\}} (t - g_n) \tag{1}$$

$$e_t = \begin{cases} \sin\left(2\pi \dfrac{t - g_k}{g_{k+1} - g_k} + \phi\right) & v_t = 1 \\ 0 & v_t = 0 \end{cases} \tag{2}$$

本研究では，学習時には声門閉鎖点を，推論時には基本周波数をピッチ情報として入力している。推論時は基本周波数系列 $\mathbf{F_0} = [F_{0,1}, \cdots, F_{0,K}]$ を用いて以下の式で励起信号を生成する。

$$e_t = \begin{cases} \sin\left(\sum_{k=1}^{t} 2\pi \dfrac{F_{0,k}}{F_s}\right) & F_{0,k} > 0 \\ 0 & F_{0,k} = 0 \end{cases} \tag{3}$$

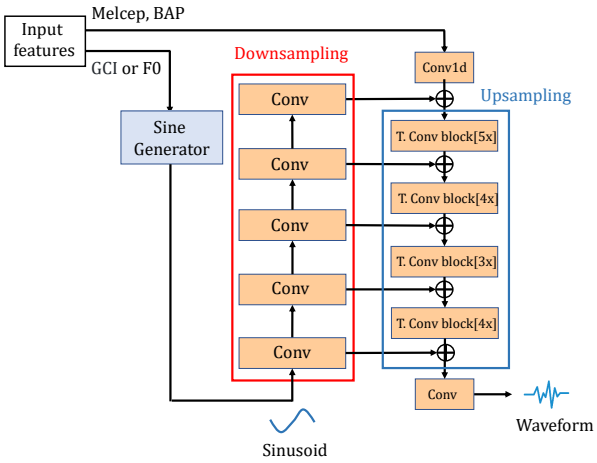


Fig. 1: Architecture of the proposed Period-HiFi-GAN generator.

4　実験

4.1　実験条件

提案法の性能を評価するため，サンプリング周波数 24 kHz の音声を用いた分析合成での客観評価および主観評価を行った。比較対象には，WORLD [3]，HiFi-GAN 及び uSFGAN [9] を用いた。データセットは JVS コーパス [15] より 100 名の日本人話者による音声のうち，95 名の 12,447 文を学習に，3 名の 90 文の音声を検証に，男女 1 名づつの 60 文の音声を評価に用いた。つまり，文献 [13] と同様，評価セットは学習セットおよび検証セットには含まれていないため，未知話者に対する分析合成を評価した。また入力の基本周波数を 0.5 倍および 1.5 倍にスケーリングして合成した場合の実験を行い，基本周波数の制御性能を評価した。

HiFi-GAN は文献 [7] による公式実装の内，モデルサイズの大きい V1 と，モデルサイズの小さい V2 を使用した。入力特徴量には 50 次元メルケプストラム，3 次元非周期性指標および対数連続 F_0 を用いた。特徴量抽出には WORLD [3] を用いて，窓長とフレームシフトを 42.7 ms と 10 ms に設定して抽出を行った。また HiFi-GAN では 80 次元メルスペクトログラム特徴量を用いた場合の比較も行った。本検討では，文献 [13] と同様，フレームシフトが 10 ms で 240 倍のアップサンプリングとなるため，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。uSFGAN は文献 [9] の公式実装を用いて，入力特徴量には上記の WORLD 特徴量を使用した。

Period-HiFi-GAN の実装には，文献 [7] の公式実装の生成器に対して励起信号入力層を追加する形で行った。励起信号入力層の各畳み込み層は，アップサ

Table 1: Results of objective evaluation without scaling of F_0.

Model	RMSE	MCD	RTF
WORLD	23.9	3.72	-
HiFi-GAN V1 (melspc)	24.3	4.68	0.29
HiFi-GAN V1 (WORLD)	23.8	4.03	0.29
HiFi-GAN V2 (melspc)	25.3	5.09	0.06
HiFi-GAN V2 (WORLD)	25.8	5.02	0.06
uSFGAN	20.5	3.58	5.85
Period-HiFi-GAN V1	24.7	3.86	0.30
Period-HiFi-GAN V2	25.8	4.67	0.05

ンプリング層の各転置畳み込み層と同一のパラメータ数で設定した。識別機に関しては HiFi-GAN で用いられているものと同一のものを使用した。入力特徴量には 50 次元メルケプストラム，3 次元非周期性指標および声門閉鎖点を用いた。声門閉鎖点の抽出には REAPER を用いた。[1] また推論時には声門閉鎖点の代わりに WORLD で抽出した線形 F_0 を用いた。

4.2 実験結果

4.2.1 客観評価実験結果

客観評価として，基本周波数の平均平方自乗誤差 (RMSE:[Hz])，メルケプストラム歪み (MCD:[dB]) およびリアルタイムファクター (RTF) を計測した。RTF の計測には Intel Xeon 6152 CPU1 コアを用いた。Table 1 に客観評価実験の結果を示す。RMSE および MCD においては uSFGAN がもっと高い品質を示し，提案法は HiFi-GAN と同等の品質となった。RTF においては，提案法は HiFi-GAN と同様の性能を示し，励起信号入力層の導入による合成速度の低下は見られず，従来法の uSFGAN よりも高速な高速生成を実現できることが確認できる。

Fig. 3 に各手法による合成音の F_0 の軌跡をプロットしたものを示す。HiFi-GAN では学習データが少ない F0 の部分での劣化が見られるが，Period-HiFi-GAN では WORLD や uSFGAN に近い軌跡となり，F0 の変化に対する忠実度が向上していることが確認できた。

4.2.2 分析合成の主観評価結果

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 20 人の成人日本語母語話者で，合計 440 文をヘッドホン聴取により評価した。

Fig. 2 に分析合成及び F0 をスケーリングした条件での主観評価実験の結果を示す。通常の分析合成の場合，男性音声の合成において提案法が最も高い品質を達成した。また F0 を 0.5 倍にした場合と男性音声で 1.5 倍にした場合でも提案法が高い品質を示した。これらの結果より，励起信号入力層の導入が低い基本周波数を持つ音声の合成に対して有効に働くことが分かった。しかし，女性音声の通常合成の場合では品質の劣化が見られた。この課題については通常の分析合成においてメルスペクトログラム条件と WORLD 特徴量条件とで品質に差があることから，HiFi-GAN の生成器自体を WORLD 特徴量用にチューニングすることで品質が改善される可能性がある。また女性音声の F0 を 1.5 倍にした場合では HiFi-GAN に大きな劣化が見られ，Period-HiFi-GAN では若干の改善が見られたものの十分な品質には至らなかった。これら課題の詳細な調査は今後の課題とする。

5 おわりに

本研究では基本周波数を制御可能な高速かつ高品質ニューラルボコーダの実現のため，HiFi-GAN に新たに励起信号を入力するネットワークを導入した Period-HiFi-GAN を提案した。実験結果より，提案法が男性音声の合成及び低い基本周波数の合成において従来法を上回る品質を達成した。今後は本研究で十分な品質が得られなかった高い基本周波数の音声について，それらの品質を改善する手法を検討する。

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.

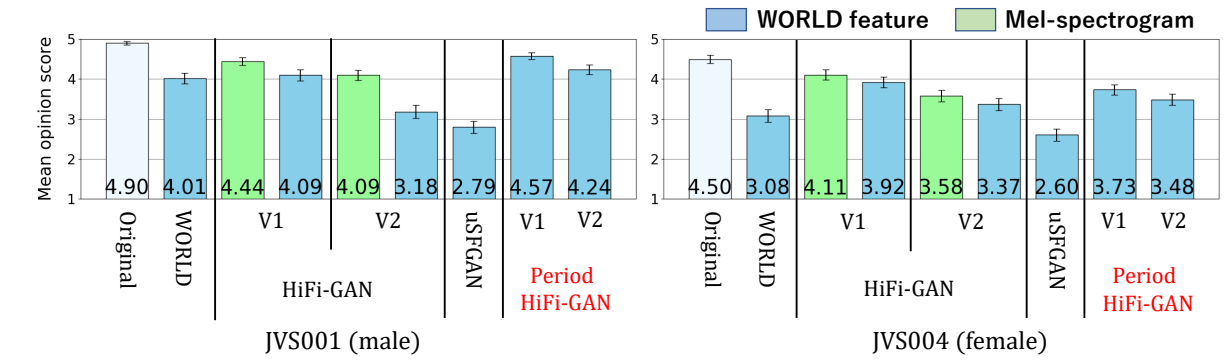
[2] A. Tamamori *et al.*, “Speaker-dependent WaveNet vocoder,” in *Proc. Interspeech*, Aug. 2017, pp. 1118–1122.

[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.

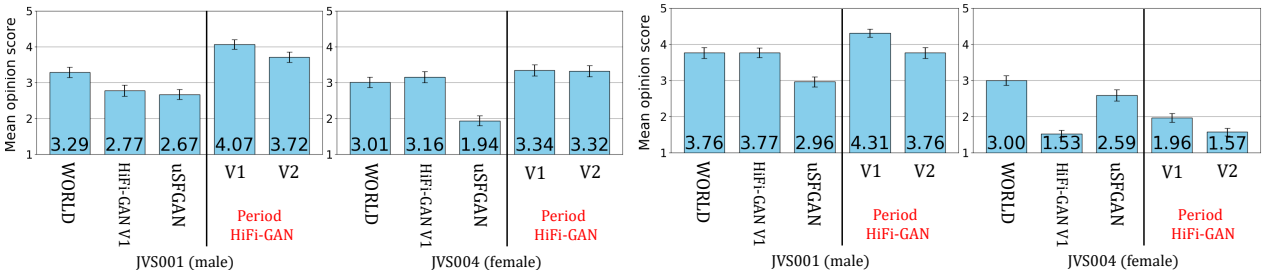
[4] I. Goodfellow *et al.*, “Generative adversarial nets,” in *Proc NIPS.*, Dec. 2014, pp. 2672–2680.

[5] R. Yamamoto *et al.*, “Parallel WaveGAN:a fast Waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in *Proc. ICASSP*, May 2020, pp. 6199–6203.

[1] https://github.com/google/REAPER



(a) $1.0 \times F_0$ condition.



(b) $0.5 \times F_0$ condition.

(c) $1.5 \times F_0$ condition.

Fig. 2: Results of the MOS test. Confidence level of the error bars was 95 %.

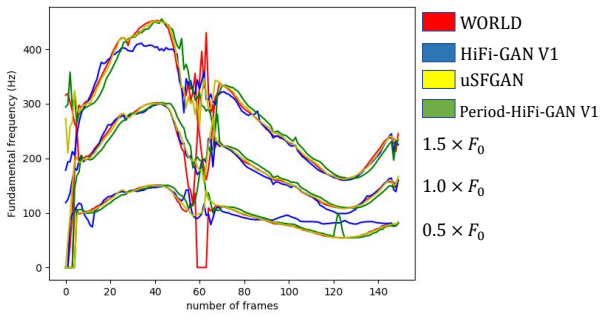


Fig. 3: Examples of F0 trajectories of male speech.

[6] K. Kumar *et al.*, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. NeurIPS*, Dec. 2019, pp. 14910–14921.

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

[8] X. Wang *et al.*, "Neural source-filter waveform models for statistical parametric speceh synthesis," *IEEE/ACM Trans. Audio, Speech, Lang, Process.*, vol. 28, pp. 402–415, 2020.

[9] R. Yoneyama *et al.*, "Unified-Source-Filter GAN: Unified source-filter network based on factorization of Quasi-Periodic Parallel WaveGAN," in *Proc. Interspeech*, Aug. 2021, pp. 2187–2191.

[10] Y. Hono *et al.*, "PeriodNet: A Non-Autoregressive Raw Waveform Generative Model With a Structure Separating Periodic and Aperiodic Components," *IEEE Access*, vol. 9, pp. 137599–137612, 2021.

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

[12] T. Okamoto *et al.*, "Multi-stream HiFi-GAN with data-driven waveform decomposition," in *Proc. ASRU*, Dec. 2021, pp. 610–617.

[13] K. Matsubara *et al.*, "Comparison of real-time multi-speaker neural vocoders on CPUs," *Acoust. Sci. Tech.* (accepted, in press).

[14] J. Valin *et al.*, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, May 2019, pp. 5891–5895.

[15] S. Takamichi *et al.*, "JSUT and JVS: free Japanese voice corpora for accelerating speech synthesis research," *Acoust. Sci. Tech*, vol. 41, pp. 761–768, Sept. 2020.