基本周波数制御可能なメルスペクトログラム入力型 HiFi-GAN の初期検討*

☆清水聡太[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

深層学習 (ニューラルネットワーク) による音声合成技術の発展に伴い，テキスト音声合成 (Text-to-Speech: TTS) や声質変換 (Voice conversion: VC) において自然音声に近い高品質な音声を合成できるようになっている [1, 2]。近年，波形生成器であるボコーダへ深層学習を導入したニューラルボコーダとして，様々なモデルが提案されている。WaveNet [3] をはじめとするニューラルボコーダは，従来のソースフィルタボコーダ [4] における品質を大きく上回り，音声合成技術の発展に大きく貢献している。

WaveNet 等の自己回帰モデル型ニューラルボコーダは合成速度が遅いという問題に対して，Parallel WaveGAN [5] や HiFi-GAN [6] など，敵対的生成ネットワーク (Generative Adversarial Network: GAN) を用いて，複数のサンプルを同時に生成することで，自己回帰構造を持たないリアルタイム合成可能なニューラルボコーダが近年研究されている。

しかし，これらのボコーダはデータ駆動型のため，学習データの範囲から大きく外れた音声を合成しようとすると品質が劣化してしまう。基本周波数 (F_0) に対する忠実度および制御性能においては，従来のソースフィルタボコーダに劣っているという問題点があった。ソースフィルタボコーダと同様に F_0 に対応した励起信号を入力する手法も提案されている [7, 8] が，これらの手法はリアルタイム合成のためにハイエンドな GPU が必要であった。このような問題を解決するため，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 に対して，F_0 に対応する励起信号およびその調波成分を入力するネットワークと，HiFi-GAN のようなアップサンプリング型モデルに対応したピッチ依存型拡張畳み込みネットワーク (Layerwise-Pitch-dependent dilated convolutional neural network: LW-PDCNN) を導入した，Harmonic-Net+ [9, 10] が提案されている。HiFi-GAN に対して，ソースフィルタボコーダのアプローチと F_0 の変化をモデル構造に組み込むことで F_0 の制御性能を改善し，CPU1 コアでのリアルタイム合成を可能にした。しかし，元の HiFi-GAN に比べると合成速度が 2 倍近く低下することが報告されている [10]。

これまで述べた F_0 制御性能の高いニューラルボコーダは，F_0 制御を行う際に音響特徴量ベースでの変換を用いるため，F_0 を含む WORLD 特徴量などが必要であった。しかし高精度な F_0 抽出には時間がかかるという問題 [11, 12] や，F_0 抽出には誤りを伴う場合があり，品質に影響を与えることがある。そこで本研究では，Harmonic-Net+で合成した F_0 制御音声を用いてデータ拡張を行い，通常の HiFi-GAN を Student model として学習することで，Harmonic-Net+の F_0 制御性能を維持しつつ合成速度の向上を図る。さらに，入力特徴量をメルスペクトログラムと F_0 制御倍率にすることで，F_0 抽出を行わず F_0 制御を可能にする試みを検討する。また，近年提案された [8] の改良版である HN-uSFGAN [13] との比較も行う。

2 Harmonic-Net+

Harmonic-Net+は F_0 制御性能の品質改善を目的として大きく 2 つの改良を加えている。1 つ目はソースフィルタボコーダの構造に倣って，HiFi-GAN の生成器に対し励起信号およびその調波成分を入力するネットワークを導入している。入力特徴量にはメルケプストラム，非周期性指標および声門閉鎖点 (学習時) または F_0(生成時) を用いる [14]。声門閉鎖点は励起信号生成器に入力され，教師音声信号の基本周波数に対応した励起信号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，入力の励起信号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。

2 つ目は基本周波数の制御性能を改善させるために HiFi-GAN のようなアップサンプリング型モデルに対応した LW-PDCNN を導入している。音声信号は一般的に周期的な信号が連続したものであるため，サンプル間で長期的な依存関係が見られる。LW-PDCNN では依存関係をもつサンプルの間隔の広さが音声の基本周波数に反比例すると仮定し，時刻ごとの基本周波数の値によって拡張サイズを変更することにより，

*Initial investigation of fundamental frequency controllable HiFi-GAN conditioned on mel-spectrograms. by SHIMIZU, Sota[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)

学習データ範囲外の基本周波数を持つ音声波形を外装可能である [7, 8, 10]。

3 検討手法

本研究では Harmonic-Net+で合成した F_0 制御音声を用いてデータ拡張を行う。Fig. 1 に検討手法の概要を示す。学習時は F_0 制御倍率を 0.05 刻みで 0.5 から 2.0 の間でランダムに決定し，Teacher model である Harmonic-Net+で合成した F_0 制御音声を教師データとして用いる。Student model である HiFi-GAN はメルスペクトログラムと F_0 制御倍率を結合したものを入力としており，F_0 抽出を行わず F_0 制御を可能にすることを期待する。

4 実験

4.1 実験条件

検討手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 [4], HiFi-GAN，HN-uSFGAN [13] および Harmonic-Net+を用いた。データセットは JSUT コーパス [15] より，日本人女性話者 1 名による 7,696 文の音声と，JVS コーパス [15] 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 96 名の 12,477 文を学習に用い，残りの 4 名の 120 文を評価に用いた。

入力特徴量には WORLD 特徴量と 80 次元メルスペクトログラムの 2 種類を用いた。WORLD 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 を用いて，いずれも WORLD 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 Harmonic-Net+の学習時には対数連続 F_0 の代わりに，REAPER[1]で抽出した声門閉鎖点を用いた。メルスペクトログラムの抽出も WORLD 特徴量と同様に窓長とフレームシフトを 42.7ms と 10ms に設定して行った。

HiFi-GAN は文献 [6] による公式実装の内，モデルサイズの大きい V1 を使用した。入力特徴量には 80 次元メルスペクトログラムを用いた。また，F_0 をスケーリングした条件での合成においては WORLD 特徴量を用いた。本検討では，フレームシフトが 10ms で 240 倍のアップサンプリングが必要となるため，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。

HN-uSFGAN は文献 [13] による公式実装を用いた。入力特徴量には WORLD 特徴量を用いた。

Harmonic-Net+の実装は，HiFi-GAN の公式実装に対して励起信号入力層と LW-PDCNN を追加した。モデルサイズは公式実装の内，V1，V2 を使用した。さらに V1 からチャネル数を少し減らした V1.5 も使用した。入力特徴量には WORLD 特徴量を用いた。励起信号入力層の各畳み込み層は，アップサ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。LW-PDCNN の実装には文献 [8] の実装を元に，アップサンプリングレートに応じてカーネルサイズと拡張サイズを変更する実装を加えた。また励起信号として sin 波の 1 から 5 倍音成分を用いた。

検討手法の実装は，HiFi-GAN の公式実装の内，モデルサイズの大きい V1 を使用した。入力特徴量には 80 次元メルスペクトログラムと F_0 制御倍率を用いた。学習時は F_0 制御倍率を 0.05 刻みで 0.5 から 2.0 の間でランダムに決定し，Harmonic-Net+で合成した F_0 制御音声を教師データとして用いた。また F_0 制御倍率が 1.0 の際は合成音声ではなく，自然音声を教師データとして用いた。

4.2 実験結果

4.2.1 客観評価実験

客観評価として，音質評価指標の WARP-Q [16] およびリアルタイムファクター (RTF) を計測した。WARP-Q はニューラル波形生成モデルの評価手法の 1 つであり，音質評価における平均オピニオン評点 (MOS) との相関性が確認されており，WARP-Q の値が小さければ小さいほど MOS 値が良くなる傾向にある。RTF の計測には Intel Xeon 6152 CPU (1 コア) を用いた。Table 1 に客観評価の結果を示す。WARP-Q においては複数話者，単一話者ともに HN-uSFGAN がもっとも高い品質となった。検討手法は複数話者において品質が大きく劣化しているが，単一話者においては品質が少し改善されている。また複数話者においては話者性が大きく失われていた。これは，入力のメルスペクトログラムに F_0 とスペクトル包絡の情報が含まれているため，F_0 だけでなく声質も変化してしまうと考えられ，今後の課題である。RTF においては，Harmonic-Net+は HN-uSFGAN と比べ高速な合成が可能であった。検討手法は HiFi-GAN と同等の性能を示し，従来法の HN-uSFGAN や Harmonic-Net+よりも高速な合成を実現している。

Fig. 2 に各手法の F_0 の平均平方二乗誤差 (F_0-RMSE: [Hz]) を F_0 制御倍率ごとにプロットしたも

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

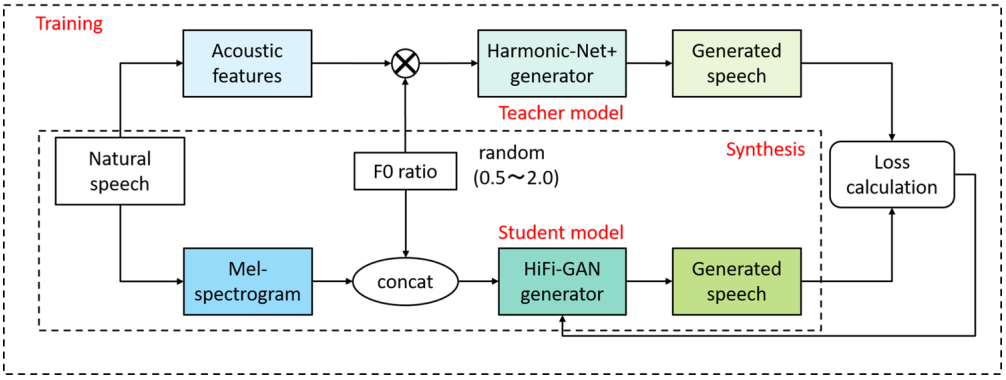


Fig. 1　Proposed method.

Table 1　Results of objective evaluation without scaling of F_0 (A: WARP-Q in JVS corpus, B: WARP-Q in JSUT corpus C: RTF).

Model	A	B	C
WORLD	0.91	0.95	-
HiFi-GAN	0.91	0.80	0.31
HN-uSFGAN	0.63	0.64	3.67
Harm+ (V1)	0.83	0.88	0.65
Harm+ (V1.5)	1.00	0.96	0.29
Harm+ (V2)	1.21	1.15	0.13
Proposed	1.87	1.07	0.32

のを示す。HiFi-GAN では制御倍率が 1.0 から離れるほど F_0-RMSE が高くなり，F_0 の変化に対する頑健性が低いことが確認できる。検討手法は従来手法と比べると F_0-RMSE が高くなっており，F_0 が少し高くなって合成されていることが確認できた。また，制御倍率が 1.0 の場合，自然音声を教師データとしているため F_0-RMSE が低くなっていると考えられる。

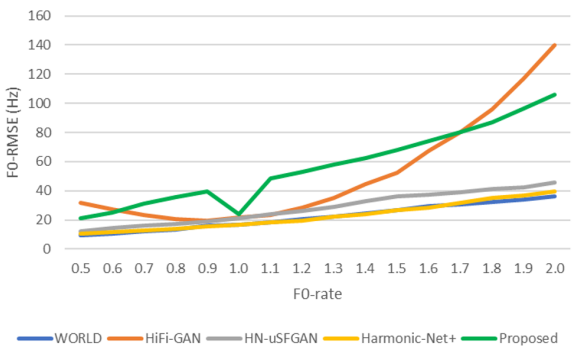


Fig. 2　Result of F0-RMSE in JSUT corpus.

4.2.2　主観評価実験

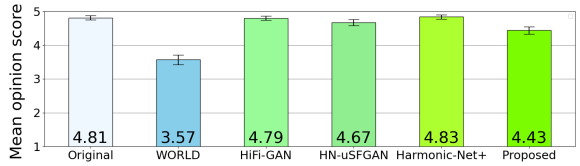
主観評価として，聴取実験による平均オピニオン評価テストを行った。単一話者モデルを使用し、F_0 制御倍率は 1.0，0.5，1.5 および 2.0 倍を用いた。F_0 制御倍率が 1.0 倍では，原音，WORLD，HiFi-GAN，HN-uSFGAN，Harmonic-Net+，および検討手法の 6 モデルを，それ以外の F_0 制御倍率では，原音と HiFi-GAN を除く 4 モデルを用いて，各条件それぞれ 20 文，合計 360 文を用いた。被験者は 10 人でヘッドホン聴取により評価した。Fig. 3 に単一話者モデルでの分析合成及び F_0 をスケーリングした条件での主観評価実験の結果を示す。Harmonic-Net+は F_0 を 2.0 倍にした場合を除いてもっとも高い品質を示した。通常の分析合成の場合，検討手法は他の従来手法には少し劣るが，高い品質を達成した。また F_0 を 0.5 倍にした場合でも WORLD を上回る高い品質をを示した。しかし，F_0 を 1.5 倍および 2.0 倍にした場合では WORLD よりも品質の劣化が見られ，今後の課題である。

5　おわりに

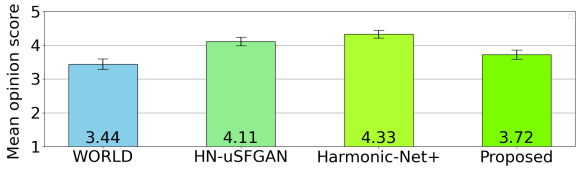
本研究では F_0 を制御可能な高速かつ高品質なニューラルボコーダの実現のため，Harmonic-Net+で合成した F_0 制御音声を用いてデータ拡張を行う，F_0 制御可能なメルスペクトログラム入力型 HiFi-GAN を検討した。実験結果より，検討手法が入力に F_0 を用いることなく F_0 を制御できることは確認できたが，指定した通りの倍率にはなっていない，話者性が損なわれてしまう点が確認された。これらは今後の課題とする。

参考文献

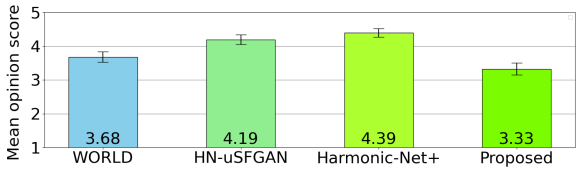
[1] J. Shen *et al.*, “Neural TTS synthesis by conditioning WavaNet on mel spectrogram predic-



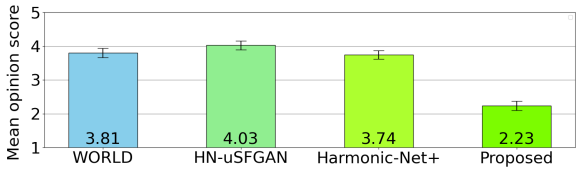
(a) 1.0 × F_0 condition.



(b) 0.5 × F_0 condition.



(c) 1.5 × F_0 condition.



(d) 2.0 × F_0 condition.

Fig. 3 Results of the MOS test in JSUT corpus. Confidence level of the error bars was 95 %.

tions," in *Proc. ICASSP*, Apr. 2018, pp. 4779–4783.

[2] 岡本，"ニューラルネットワークに基づく音声波形生成モデル"，音響誌，vol. 78，no. 6，pp. 328–337，June 2022.

[3] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," in *Proc. SSW9*, Sept. 2016, p. 125.

[4] M. Morise, F. Yokomori, and K. Ozawa, "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, July 2016.

[5] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, May 2020, pp. 6199–6203.

[6] J. Kong *et al.*, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Proc. NeurIPS*, Dec. 2020, pp. 17 022–17 033.

[7] X. Wang *et al.*, "Neural source-filter waveform models for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang, Process*, vol. 28, pp. 402–415, Nov. 2019.

[8] R. Yoneyama *et al.*, "Unified Source-Filter GAN: Unified Source-Filter Network Based On Factorization of Quasi-Periodic Parallel WaveGAN," in *Proc. Interspeech*, Sept. 2021, pp. 2187–2191.

[9] 松原ら，"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"，音講論，pp. 901–904，Mar. 2022.

[10] ——，"Harmonic-Net+：高調波入力とLayerwise-Quasi-Periodic 畳み込みを用いた基本周波数制御可能な高速ニューラルボコーダ"，音講論，Sept. 2022.

[11] H. Kawahara *et al.*, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," in *Proc. Interspeech*, Sept. 2005, pp. 537–540.

[12] M. Morise *et al.*, "Harvest: A high-performance fundamental frequency estimator from speech signals," in *Proc. Interspeech*, Aug. 2017, pp. 2321–2325.

[13] R. Yoneyama *et al.*, "Unified Source-Filter GAN with Harmonic-plus-Noise Source Excitation Generation," in *Proc. Interspeech*, 2022.

[14] 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. 137 599–137 612, Oct. 2021.

[15] 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.

[16] W. A. Jassim *et al.*, "WARP-Q: Quality prediction for generative neural speech codecs," in *Proc. ICASSP*, June 2021, pp. 401–405.