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 *

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

1 はじめに

ニューラルネットに基づく音声合成は，近年の大幅な研究の進展により，CPU のみでもリアルタイムに高品質な合成が可能となっており [1,2]，テキスト音声合成においては音素系列から高品質な音声波形を 1 つのニューラルネットで直接生成可能な End-to-end モデルの検討も進んでいる [3,4]。

このように，ニューラルネットワークに基づく方式は音声コーパスに対してデータ駆動型であるため高品質な生成が可能であるが，それ故に基本周波数 (f_{o} [5]) 等の制御においては学習データに存在しない範囲外の f_{o} を含む音声波形の生成は難しく，制御性能は従来の信号処理に基づくソースフィルタ型ボコーダ (e.g. WORLD [6]) には及ばない。

ニューラル波形生成モデルにおける f_{o} の制御性能を向上させるリアルタイム方式として，f_{o} に対応する高調波を入力とする Neural source filter (NSF) [7]，f_{o} に応じて拡張畳み込みの受容野サイズを動的に変化させる Quasi-Periodic Parallel WaveGAN (QPPWG) [8]，NSF と QPPWG を統合し，高精度化した Unifiled source-filter GAN (uSFGAN) [9]，周期成分と非周期成分を別々のネットワークとし，周期成分には f_{o} に対応したサイン波，非周期成分には白色雑音を入力とする PeriodNet [10]，PeriodNet と uSFGAN を統合した HN-uSFGAN [11] 等が提案されている。これらのモデルは単純な WORLD 特徴量を入力とし f_{o} の倍率を変化させた場合のモデルよりも高精度に f_{o} を制御しつつ，WORLD よりも高品質な合成を実現できる。しかし，これらのモデルはリアルタイム生成にはハイエンドな GPU が必要であるという課題が残る。また，PeriodNet は歌声生成では高品質であるが，通常発話の生成は品質が極めて劣化する問題がある [12]。

f_{o} の制御性能を維持しつつ，高品質かつ CPU のみでもリアルタイム生成可能なニューラル音声波形生成モデルとして，我々は Period-HiFi-GAN を提案した [13]。このモデルは，高速かつ高品質な音声波形生成が可能な HiFi-GAN の生成器に f_{o} に対応したサイン波を入力可能なダウンサンプリングネットワークを導入することにより実現される。複数話者コーパスを用いた未知話者合成の実験において，男性音声については f_{o} が 0.5 倍から 1.5 倍において WORLD や uSFGAN よりも高品質な生成を実現できるが，女性音声については f_{o} が 1.0 倍や 1.5 倍においては従来方式に品質が及ばないという課題があった [13]。

そこで，女性音声の合成品質を向上させるために，ダウンサンプリングネットワークへの入力を単純なサイン波ではなく高調波へと拡張させた Harmonic-Net，さらに QPPWG や uSFGAN で用いられている f_{o} に応じて拡張畳み込みの受容野サイズを動的に変化させる Quasi-Periodic 畳み込み層 (Pitch-dependent dilated CNN: PDCNN) をアップサンプリング型の HiFi-GAN 生成器に導入した Harmonic-Net+を提案する。

2 提案法

2.1 Harmonic-Net

Harmonic-Net は，PeriodNet や Period-HiFi-GAN と同様，学習時は声門閉鎖点 (GCI) の周期に対応するサイン波およびその高調波を，推論時は f_{o} に対応するサイン波およびその高調波をダウンサンプリングネットワークへ入力する。アップサンプリングネットワークには，Period-HiFi-GAN と同様，WORLD 特徴量のうちメルケプストラム (melcep) および非周期性指標 (BAP) のみを入力とする (Fig. 1)。声門閉鎖点の秒数にサンプリング周波数 f_{s} を掛けた系列を $\boldsymbol{g} = [g_1, \cdots, g_q, \cdots, g_Q]$ とし，各時刻における有声/無声系列を $\boldsymbol{v} = [v_1, \cdots, v_t, \cdots, v_T]$ とすると，学習時における i 次高調波に対応する励起信号 $e_{t,i}$ は，

$$l = \arg \min_{\{q:g_q<t\}} (t - g_q) \tag{1}$$

$$e_{t,i} = \begin{cases} \sin\left(2\pi i \dfrac{t-g_l}{g_{l+1}-g_l} + \phi\right) & v_t = 1 \\ 0 & v_t = 0 \end{cases} \tag{2}$$

となる。ここで，ϕ は時刻 t における初期位相である。一方，推論時の $e_{t,i}$ は，f_{o} 系列を $\boldsymbol{f}_{\mathrm{o}}$ =

---

* Harmonic-Net+: Fundamental frequency controllable fast neural vocoder with harmonic wave input and Layerwise-Quasi-Periodic CNNs. by MATSUBARA, Keisuke[2,1], OKAMOTO, Takuma[1], TAKASHIMA, Ryoichi[2], TAKIGUCHI, Tetsuya[2], TODA, Tomoki[3,1], KAWAI, Hisashi[1] ([1]NICT, [2]Kobe Univ, [3]Nagoya Univ)

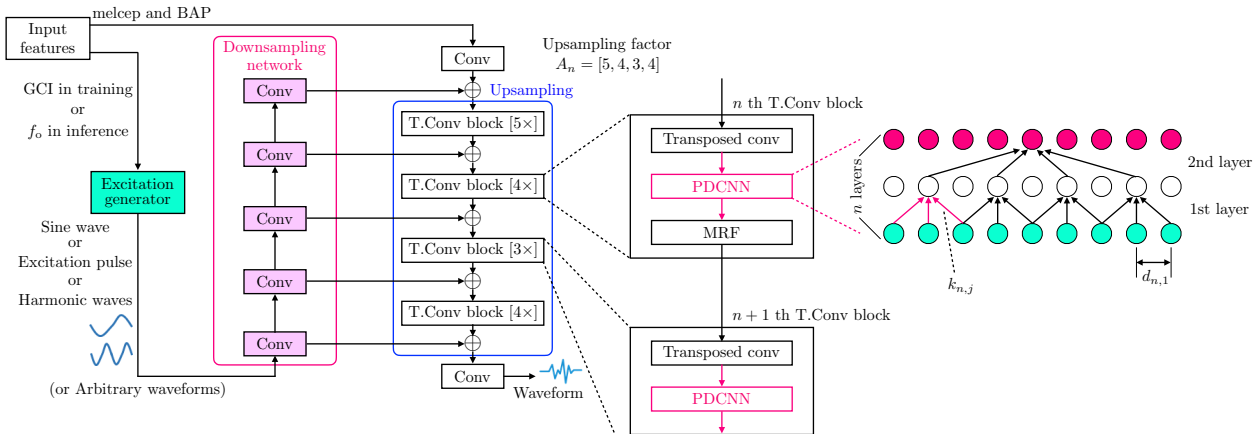


Fig. 1　Network architecture of Harmonic-Net+ generator with layerwise PDCNNs.

$[f_{o,1}, \cdots, f_{o,l}, \cdots, f_{o,L}]$ とすると，

$$e_{t,i} = \begin{cases} \sin\left(\sum_{l=1}^{t} 2\pi i \frac{f_{o,l}}{f_s}\right) & f_{o,l} > 0 \\ 0 & f_{o,l} = 0 \end{cases} \tag{3}$$

となる。Period-HiFi-GAN ではダウンサンプリングネットワークへの入力はサイン波のみ ($i = 1$) であり，NSF では高調波を入力するもののそれらの重みは固定であったが，Harmonic-Net では入力チャネルを i とすることにより，i 次高調波までがネットワークに (=学習可能な重み付けで) 入力される。推論時は式 (3) における $f_{o,l}$ の倍率を変えることにより，基本周波数の制御が可能となる。また，ダウンサンプリングネットワークへの入力は波形であれば何でもよいため，ソースフィルタ型ニューラルボコーダで用いられるパルス型励起信号や WORLD 等において合成した音声の入力も可能である。

2.2　Harmonic-Net+

ニューラル波形生成モデルで広く用いられる通常の非因果的拡張畳み込み [2] の出力 $\boldsymbol{y}_t^{(o)}$ は，

$$\boldsymbol{y}_t^{(o)} = \sum_{k=0}^{K} \left(\boldsymbol{W}^{(k)} \times \boldsymbol{y}_{t-(\frac{K}{2}-k)d}^{(i)} \right) \tag{4}$$

となる。ここで，d は受容野サイズ，K はカーネルサイズ，$\boldsymbol{W}^{(k)}$ は k 番目の学習可能な 1×1 畳み込みフィルタである。式 (4) のように，通常の拡張畳み込みは受容野サイズ d が時不変であるのに対して，PDCNN では，$f_{o,l}$ の値に応じて受容野サイズ d' を

$$E_t = \frac{f_s}{f_{o,t} \times a} \tag{5}$$

$$d' = E_t \times d \tag{6}$$

のように時間的に変化させる。ここで，$f_{o,t}$ は時刻 t における基本周波数であり，a は重み係数 (ハイパーパラメータ) である [8]。QPPWG や uSFGAN では，$f_{o,t}$ に依存した PDCNN を用いることにより，学習データ範囲外の基本周波数の外装を可能としている [8,9]。

QPPWG や uSFGAN は入力する音響特徴量を最初に音声波形のサンプリング周波数までアップサンプリングした上で条件付けするパラレル生成モデルであるのに対して [2]，HiFi-GAN 生成器は音響特徴量のみを徐々にアップサンプリングして音声波形を直接生成するアップサンプリング型モデルであるため [2]，Harmonic-Net には PDCNN を直接導入することができない。そこで，HiFi-GAN 生成器のようなアップサンプリング型モデルに PDCNN を適用するために，Layerwise-Quasi-Periodic 畳み込み (Layerwise-PDCNN) を提案する。Fig. 1 における n 番目の T.Conv ブロックの時間解像度を $F_{s,n}$ とし，j 層目の PDCNN のカーネルサイズを $k_{n,j}$ とすると，対応する受容野サイズ $d_{n,j}$ は，

$$f_{s,n} = f_{s,0} \prod_{m=1}^{n} A_m \tag{7}$$

$$k_{n,j} = \begin{cases} A_j & j \neq 1 \\ 3 & j = 1 \end{cases} \tag{8}$$

$$d_{n,j} = \begin{cases} d_{n,2} \prod_{m=2}^{j-1} A_m & j > 2 \\ 2d_{n,1} & j = 2 \\ \frac{f_{s,n}}{f_{o,t} \times a} & j = 1 \end{cases} \tag{9}$$

となる。ここで，$f_{s,0}$ は入力音響特徴量の時間解像度，A_n は n 番目の T.Conv ブロックのアップサンプリング率である。1 層目の Layerwise-PDCNN のカーネルサイズと受容野サイズは従来の PDCNN と同様に設計され，2 層目以降はアップサンプリング率に従って定義される。これにより，アップサンプリング型音声波形生成モデルにも f_o に応じた拡張畳み込みが可能となる。

3 実験

3.1 実験条件

提案法の有効性を確認するために，JVS コーパス [14] を用いた複数話者モデル ($f_s = 24$ kHz，jvs005 から jvs1000 のそれぞれ 130 文を学習セット，jvs001 から jvs004 を評価セット) における未知話者合成および東北きりたんコーパス [15] を用いたフル帯域歌声合成モデル ($f_s = 48$ kHz，学習および評価条件は文献 [12] と同様) を学習した。Harmonic-Net および Harmonic-Net+は Period-HiFi-GAN のモデルを拡張し，式 (9) における a は 4.0 とした。予備検討の結果，$i = 5$ 次までの高調波成分を入力とした。複数話者モデルは WORLD，uSFGAN，複数話者自己回帰型 WaveNet ボコーダ [16]，HiFi-GAN およびそのメルスペクトログラム入力 (melspc) と比較した。フル帯域歌声合成モデルでは，Harmonic-Net+の非周期成分に劣化が見られる課題があったため，Harmonic-Net のみの評価とし，WORLD，HiFi-GAN，HiFi-GAN (melspc) および PeriodNet と比較した。QPPWG や Period-HiFi-GAN と同様，f_o は 1.0 倍，0.5 倍および 1.5 倍の条件を評価した。複数話者モデルにおいては，男性音声が 0.5 倍時に，女性音声が 1.5 倍時にそれぞれ学習データの範囲外であった。

提案法の合成速度をリアルタイムファクター (RTF) として Intel Xeon 6152 CPU (1 コア) を用いて計測した。また，合成音声の音質評価のための聴取実験を行った。実験参加者は正常な聴力を有する成人 20 名とし，ヘッドホンにより原音および合成音声を聴取した。

3.2 実験結果

複数話者モデルの RTF は，HiFi-GAN，Harmonic-Net および Harmonic-Net+においてそれぞれ 0.38，0.40 および 0.66 であり，フル帯域歌声合成モデルではそれぞれ 0.86，0.91 および 1.56 であった。Layerwise-PDCNN を導入することにより合成速度が遅くなってしまう課題があるが，複数話者モデルではリアルタイム生成が可能であることがわかる。

複数話者モデルの聴取実験の平均オピニオン評点 (MOS) の結果を Fig. 2，フル帯域歌声合成モデルの結果を Fig. 3 にそれぞれ示す。複数話者モデルの結果では，女性音声 $1.0 \times f_o$ は自己回帰型 WaveNet(GPU を用いてもリアルタイム合成不可) の方が優れているが，それ以外では Harmonic-Net+が最も高品質な合成を達成しており，特に女性音声 $1.0 \times f_o$ では Layerwise-PDCNN の効果が確認できる。また，フル帯域歌声合成モデルの結果においても，Harmonic-Net は PeriodNet(リアルタイム生成には GPU が必要) と同等またはそれ以上の高音質を実現している。

以上により，提案法の有効性が確認できる。

4 おわりに

HiFi-GAN 生成器に高調波入力型ダウンサンプリングネットワークを導入した Harmonic-Net およびさらに f_o に対応した Layerwise-Quasi-Periodic 畳み込みを導入した Harmonic-Net+を提案した。聴取実験により提案法の有効性を確認した。

参考文献

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

[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] J. Kim *et al.*, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *Proc. ICML*, July 2021, pp. 5530–5540.

[4] D. Lim *et al.*, “JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech,” in *Proc. Interspeech*, Sept. 2022.

[5] I. R. Titze *et al.*, “Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization,” *J. Acoust. Soc. Am.*, vol. 137, no. 5, pp. 3005–3007, May 2015.

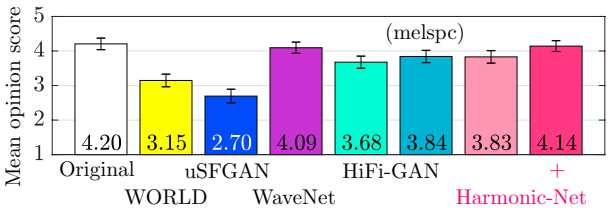
[6] M. Morise *et al.*, “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE trans. Inf. Syst.*, vol. E99-D, no. 7, pp. 1877–1884, July 2016.

[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, 2020.

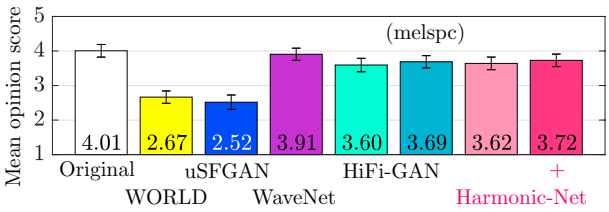
[8] Y.-C. Wu *et al.*, “Quasi-Periodic Parallel WaveGAN: A non-autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 792–806, 2021.

[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.

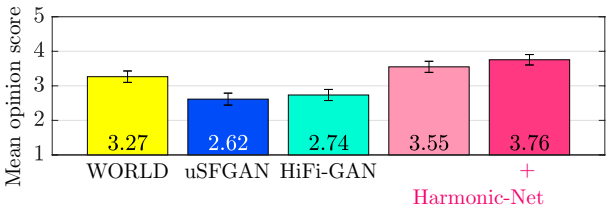
[10] Y. Hono *et al.*, “PeriodNet: A non-autoregressive raw waveform generative model



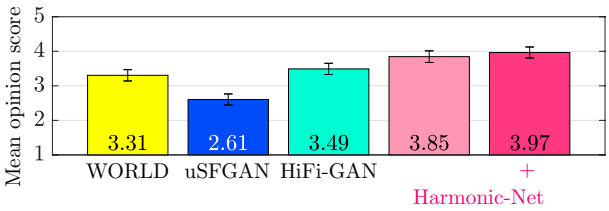
(a) Normal condition ($1.0 \times f_{\mathrm{o}}$) of male speech



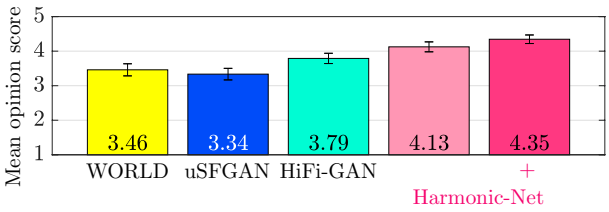
(b) Normal condition ($1.0 \times f_{\mathrm{o}}$) of female speech



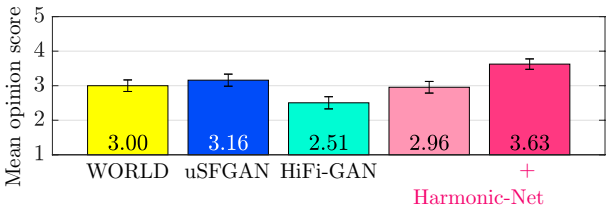
(c) $0.5 \times f_{\mathrm{o}}$ condition of male speech



(d) $0.5 \times f_{\mathrm{o}}$ condition of female speech



(e) $1.5 \times f_{\mathrm{o}}$ condition of male speech



(f) $1.5 \times f_{\mathrm{o}}$ condition of female speech

Fig. 2 Results of MOS test for unseen speaker synthesis in normal and f_{o}-conversion conditions. Confidence level of the error bars is 95%.

with a structure separating periodic and aperiodic components," *IEEE Access*, vol. 9, pp. 137599–137612, 2021.

[11] R. Yoneyama *et al.*, "Unified Source-Filter GAN with harmonic-plus-noise source excitation generation," in *Proc. Interspeech*, Sept. 2022.

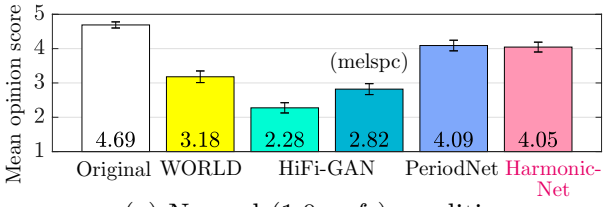
[12] K. Matsubara *et al.*, "Full-band LPCNet: A real-time neural vocoder for 48 khz audio with a CPU," *IEEE Access*, vol. 9, pp. 94923–94933, 2021.

[13] 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.

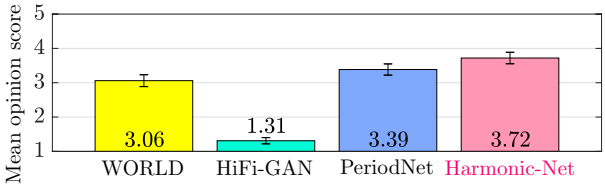
[14] S. Takamichi *et al.*, "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *Acoust. Sci. Tech.*, vol. 41, no. 5, pp. 761–768, Sept. 2020.

[15] I. Ogawa *et al.*, "Tohoku Kiritan singing database: A singing database for statistical parametric singing synthesis using Japanese pop songs," *Acoust. Sci. Tech.*, vol. 42, no. 3, pp. 140–145, May 2021.

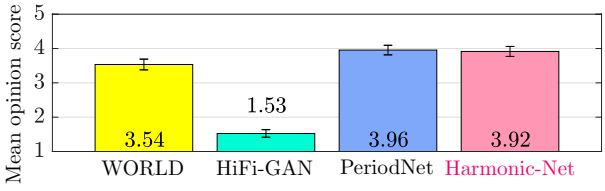
[16] T. Hayashi *et al.*, "An investigation of multi-speaker training for WaveNet vocoder," in *Proc. ASRU*, Dec. 2017, pp. 712–718.



(a) Normal ($1.0 \times f_{\mathrm{o}}$) condition



(b) $0.5 \times f_{\mathrm{o}}$ condition



(c) $1.5 \times f_{\mathrm{o}}$ condition

Fig. 3 Results of MOS test for full-band singing voice synthesis in the normal and f_{o}-conversion conditions. Confidence level of the error bars is 95%.