# オペラ歌唱音声合成におけるニューラルボコーダの比較検討*

☆清水聡太 (神戸大), 松原圭亮 (神戸大), 足立優司 (メック株式会社),
田井清登 (メック株式会社), 高島遼一 (神戸大), 滝口哲也 (神戸大)

## 1 はじめに

歌声合成技術は娯楽分野において広く普及し注目を集めている。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，歌声合成の分野においても高品質な音声の合成が可能になっている [1]。近年，ニューラルボコーダとして様々なモデルが提案されており，WaveNet [2] 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 [3, 4] における品質を大きく上回り，歌声合成技術の発展に大きく貢献している。

WaveNet は高品質な音声合成が可能だが，自己回帰構造を持ち，巨大なモデル構造であることから，合成速度が遅いという問題がある。この問題を解決するため，WaveRNN[5] や LPCNet[6] などの，自己回帰構造でありながら，モデルパラメータ数を削減することにより，高速合成を可能にするモデルが提案されている。また，MelGAN[7]，Parallel WaveGAN [8]，HiFi-GAN [9] 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リアルタイム合成可能な，自己回帰構造を持たないニューラルボコーダが近年研究されている。また，歌声合成を高品質に行えるニューラルボコーダとして PeriodNet [10] が提案されている。

本稿で対象としているオペラ歌唱は，従来の歌声合成で扱う童謡や J-POP とは，ビブラートやピッチ，母音などの特徴が異っており [11, 12, 13]，歌声合成にニューラルボコーダを用いた先行研究はあるが [1]，オペラ歌唱音声を対象としたニューラルボコーダの研究はほとんどない。そこで本稿では，オペラ歌唱音声の分析合成実験により，ニューラルボコーダの性能を比較評価する。

## 2 ニューラルボコーダ

本稿では，WaveNet，Parallel WaveGAN，PeriodNet，HiFi-GAN の 4 種類のニューラルボコーダを評価した。

### 2.1 WaveNet

WaveNet は自己回帰構造をもつ畳み込みニューラルネットワークである。過去の音声サンプルを条件として次のサンプルを予測することで，音声波形を推定する。WaveNet では予測を簡単にするため，$\mu$-law 変換を用いて音声波形を 16bit から 8bit に量子化している。また，波形推定を回帰問題としてではなく，256 種類の離散型分類問題としている。

WaveNet は高品質な音声を生成可能だが，自己回帰構造であることと，巨大なモデル構造であることから，合成速度が遅い点や，予測誤差による雑音成分が問題としてある [14]。

### 2.2 Parallel WaveGAN

Parallel WaveGAN は，Parallel WaveNet [15] をベースモデルとする GAN である。入力はホワイトノイズと音響特徴量であり，非自己回帰型の Generator が複数のサンプルを同時に生成する。Discriminator は Generator が生成した音声波形と実際の音声波形を判別できるように学習される。また，通常の GAN で用いられる Adversarial loss に加えて，出力波形と目標波形に短時間フーリエ変換 (STFT) を適用した上で計算される誤差である，STFT loss を導入して Generator の学習を補助している。STFT は，時間解像度と周波数解像度がトレードオフの関係にあるため，一方を細かく分析しようとすると，もう一方の分析が粗くなってしまう。そこで，解像度の異なる複数の STFT の損失を導入することで，時間解像度と周波数解像度の両方を担保している。

Parallel WaveGAN は自己回帰構造をもたないため，複数のサンプルを同時に生成することで高速合成が可能であり，リアルタイム合成を実現している。また，Parallel WaveNet と比べ，知識蒸留なしで直接並列生成モデルを学習可能である。

### 2.3 Period Net

PeriodNet は GAN をベースとする非自己回帰構造のニューラルネットワークである。PeriodNet では音声波形を周期成分と非周期成分に分けてモデル化しており，それぞれ異なる Generator によって推定を行っている。周期成分を生成する Generator には正

*A comparison of neural vocoders in opera singing voice synthesis. by Sota Shimizu (Kobe Univ.), Keisuke Matsubara (Kobe Univ.), Yuji Adachi (MEC Company Ltd.), Kiyoto Tai (MEC Company Ltd.), Ryoichi Takashima (Kobe Univ.), Tetsuya Takiguchi (Kobe Univ.)

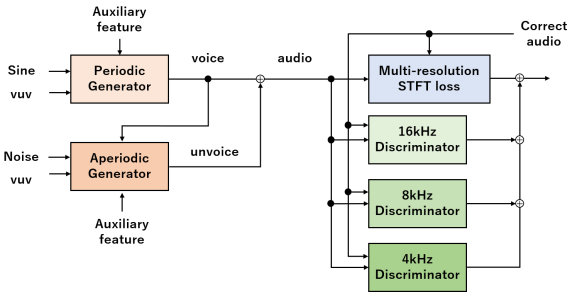


Fig. 1　PeriodNet structure

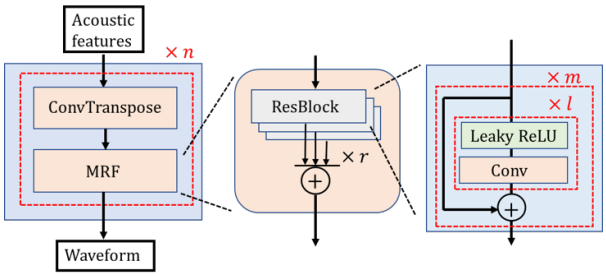


Fig. 2　Generator of HiFi-GAN structure

弦波を，非周期成分を生成する Generator にはガウスノイズをそれぞれ励起信号として入力することで，各成分を推定している。励起信号を明示的に入力しているため，学習データの範囲外の $F_0$ を持つ音声でも比較的頑健に合成が行える。Generator は Parallel WaveGAN と同様に，Adversarial loss，STFT loss で学習を補助している。Discriminator は，音声信号をダウンサンプリングし，複数のサンプリング周波数について別々に識別器を学習している。

PeriodNet は Parallel WaveGAN と同様に，自己回帰構造をもたないため，複数のサンプルを同時に生成することで高速合成が可能であり，リアルタイム合成を実現している。

### 2.4　HiFi-GAN

HiFi-GAN は GAN をベースとする非自己回帰構造のニューラルネットワークである。転置畳み込み (Transposed convolution) を用いて，入力された音響特徴量を音声波形に変換する Generator と，複数のサンプリング周波数と受容野をもつ，2 種類の Discriminator から構成される。

Generator は通常の畳み込み層と転置畳み込み層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，異なる受容野をもつ，複数の畳み込み層の出力を足し合わせることで，音声波形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。Discriminator 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，音声波形を様々な間隔でサンプリングして真偽を識別する 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音声に対して，別々の Discriminator で識別する。MPD では長さ $T$ の音声波形に対して間隔 $d$ でサンプリングを行い，$(T/d) \times d$ の 2 次元データに変形した後に Discriminator 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 Discriminator を用いて学習を行う。これらの処理により，音声波形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できる。

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であり，リアルタイム合成を実現している。

## 3　実験評価

### 3.1　実験条件

本研究ではオペラ歌唱音声を対象として，分析合成の実験によりニューラルボコーダの性能比較をした。女性歌手 1 名による日本語アカペラオペラ歌唱音声 48 曲からなる約 93 分の音声データを収録した。サンプリング周波数は 16kHz，量子化ビット数 16bit である。この 48 曲のうち，43 曲 (約 85 分) を学習データ，5 曲 (約 8 分) をテストデータに用いた。学習データ 43 曲のうち 5 曲 (約 9 分) は，Parallel WaveGAN，PeriodNet，HiFi-GAN の学習時に開発データとして用いた。モデルの学習には GeForce RTX 3090 の GPU を 1 枚使用した。

音響特徴量には，WORLD[3] によってフレームシフト 5ms で抽出したメルケプストラム 50 次元，対数基本周波数 1 次元，非周期性指標 1 次元，有声無声区間情報 1 次元を使用した。

WaveNet のモデル構造は，Dilation が 1, 2, 4, ..., 256, 512 の Dilated Causal Convolution を 3 回繰り返した 30 層の畳み込み層で，カーネルサイズは 2，バッチサイズは 1 × 50,000 とした。学習回数は 50 万回とし，学習に約 7 日かかった。また，雑音成分を抑えるため，時不変ノイズシェーピング法 [14] を適用した。

Parallel WaveGAN の Generator は，WaveNet と同様に 30 層の畳み込み層で，カーネルサイズは 3 とした。Discriminator は，カーネルサイズ 3 の畳み込み層と Leaky ReLU を交互につなげた，19 層のものを使用した。バッチサイズは 2 × 25,600，学習回数は 40 万回とし，学習に約 49 時間かかった。

PeriodNet のモデル構造は Fig. 1 のようになっている。Parallel WaveGAN と同様の Generator を，周期

Table 1　Results of objective evaluation

| Model | WORLD | WaveNet | Parallel WaveGAN | PeriodNet | HiFi-GAN |
|---|---|---|---|---|---|
| MCD [dB] | 3.49 | 3.99 | **3.39** | **3.13** | 4.05 |
| $F_0$-RMSE [Hz] | 14.23 | 13.28 | 11.51 | **11.29** | 21.14 |

成分と非周期成分の Generator でそれぞれ 2 つ，Discriminator は，サンプリング周波数が 16kHz，8kHz，4kHz の Discriminator でそれぞれ 3 つとした。バッチサイズは 2 × 25,600，学習回数は 40 万回とし，学習に約 25 時間かかった。

HiFi-GAN の Generator のモデル構造は Fig. 2 のようになっている。今回の実験では $n$=4，$r$=12，$m$=2，$l$=3 とし，ResBlock のカーネルサイズを 3, 7, 11 とした。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，今回の実験ではフレームシフトが 5ms のため，80 倍のアップサンプリングとなる。したがって，アップサンプリング数を 2, 5, 4, 2 とし，転置畳み込み層のカーネルサイズを 4, 10, 8, 4 とした。MSD は PeriodNet と同様のサンプリング周波数とし，MPD は音声波形を 2, 3, 5, 7, 11 の間隔でサンプリングした。バッチサイズは 2 × 25,600，学習回数は 10 万回とし，学習に約 6 時間かかった。

また，WaveNet，Parallel WaveGAN，PeriodNet，HiFi-GAN の 4 つのニューラルボコーダに，ソースフィルタボコーダである WORLD を加え，5 つのボコーダで実験を行った。

### 3.2　実験結果

#### 3.2.1　客観評価実験

本稿では客観評価指標としてメルケプストラム歪み (Mel-cepstrum distortion; MCD) [dB] と基本周波数の二乗誤差 ($F_0$-root mean square error; $F_0$-RMSE) [Hz] を用いた。Table 1 に客観評価実験の結果を示す。Table 1 より，Parallel WaveGAN，PeriodNet の MCD が小さいことがわかる。これは，2 つのモデルで STFT loss を導入しているため，メルケプストラムの推定精度が高かったものと考えられる。WaveNet は，予測誤差による雑音が時折発生しており，そのため MCD が大きかったと考えられる。$F_0$-RMSE に関しては，Table 1 より，特に PeriodNet が小さいことがわかる。これは，PeriodNet では明示的に励起信号を入力しているため，$F_0$ の推定精度が高いことが考えられる。WaveNet は，合成音声に音程が不安定な箇所が見られ，$F_0$-RMSE が高いという結果であった。HiFi-GAN に関しては，合成品質が悪く，MCD，$F_0$-RMSE ともに大きい値となった。

#### 3.2.2　主観評価実験

主観評価指標として平均オピニオン指標 (MOS) を用いた。1 が非常に悪い音声，5 を非常に良い音声として，5 段階評価を行った。被験者は 10 人で，テストセット 50 フレーズに対して，5 つのボコーダに収録音声を加えた計 6 条件とし，300 フレーズをヘッドホン聴取により評価した。Fig. 3 に主観評価実験の結果を示す。Fig. 3 より，WORLD が最も高いスコアを示した。オペラ歌唱音声は調波構造の強い音声であるため，音声の調波構造に強い仮定を置いているソースフィルタボコーダにとっては，発話を対象とした音声合成に比べて，高品質な合成が行えたものと考えられる。Parallel WaveGAN，PeriodNet は WOLRD に次いで高いスコアを示した。またこの 2 つのモデルに，$t$ 検定による有意差は見られなかった。WaveNet は，予測誤差による雑音が時折発生しており，そのためスコアが低かったものと考えられる。また，WaveNet の合成音声には音程が不安定な箇所があり，これもスコアが低い要因と考えられる。HiFi-GAN に関しては合成品質が悪く，スコアが低かった。

WaveNet，Parallel WaveGAN および HiFi-GAN は，元々発話音声の合成を対象に提案されたものであり，PeriodNet も一般的な歌唱音声を対象に提案されたものである。今回の検討では，各モデルのハイパーパラメータをデフォルトのものから大きく変えなかったため，特に WaveNet と HiFi-GAN については，オペラ歌唱音声の合成が高品質に行えなかったものと考えられる。そのため，ニューラルボコーダについてはハイパーパラメータの調整次第でさらに品質が改善する可能性はあるが，その調査については今後の課題とする。

## 4　おわりに

本稿ではオペラ歌唱音声を対象に，WaveNet，Parallel WaveGAN，PeriodNet，HiFi-GAN の 4 種類のニューラルボコーダを性能比較した。客観評価実験より，Parallel WaveGAN，PeriodNet はメルケプストラムの推定精度が高いことが示された。また，PeriodNet は $F_0$ の再現精度が高いことが示された。主観評価実験では，WORLD に次いで Parallel WaveGAN，

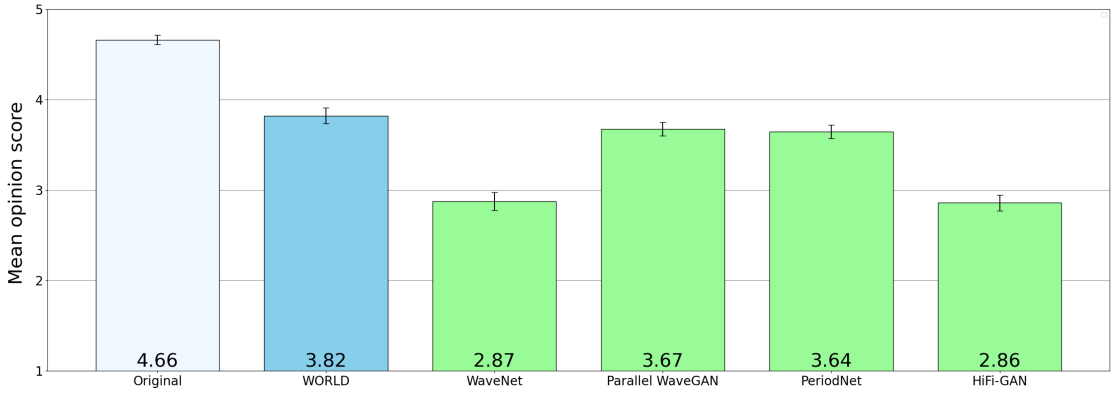


Fig. 3　MOS evaluation result

PeriodNet が高いスコアを示した。

調波構造の強い歌声音声を対象とした分析合成では，ソースフィルタボコーダである WORLD が高品質な音声合成が可能であることがわかった。しかし，歌詞付きの楽譜から歌声合成を行う場合では，ファインチューニングを行えるニューラルボコーダが高品質に音声合成をできる可能性がある。

今後の課題としては，オペラ歌唱音声に適したハイパラメータの調整による品質の向上，サンプリング周波数 48kHz の音声での実験，ニューラルボコーダを用いた歌詞付き楽譜からの歌声合成が挙げられる。

## 参考文献


[1] 和田ら，“歌声合成におけるニューラルボコーダの比較検討”，*IEICE Technical Report, SP2019-42(2019-12)*，2019，pp. 85–90.

[2] A. van den Oord *et al.*, “WaveNet: A generative model for raw audio,” *arXiv:1609.03499*, 2016.

[3] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[4] H. Kawahara *et al.*, “Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation,” in *Proc.ICASSP*, 2008, pp. 3933–3936.

[5] N. Kalchbrenner *et al.*, “Efficient Neural Audio Synthesis,” *arXiv:1802.08435*, 2018.

[6] J. S. Jean-Marc Valin, “LPCNet: Improving Neural Speech Synthesis Through Linear Prediction,” *arXiv:1810.11846*, 2019.

[7] K. Kumar *et al.*, “MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis,” *arXiv:1910.06711*, 2019.

[8] R. Yamamoto, E. Song, and J.-M. Kim, “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” *arXiv:1910.11480*, 2019.

[9] J. Kong *et al.*, “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,” *arXiv:2010.05646*, 2020.

[10] Y. Hono *et al.*, “PeriodNet: A non-autoregressive waveform generation model with a structure separating periodic and aperiodic components,” *arXiv:2102.07786v1*, 2019.

[11] Johan Sundberg *et al.*，歌声の科学．　東京電機大学出版局, 2007，pp. 165–177.

[12] 片平健太 *et al.*，“歌声の母音変化を考慮した歌声合成の検討”，日本音響学会秋季研究発表会講演論文集，2019，pp. 1007–1010.

[13] ——, “母音の発音と歌唱速度の変化を考慮したアカペラオペラ歌声合成,” in 日本音響学会春季研究発表会講演論文集, 2021, pp. 991–994.

[14] K. Tachibana *et al.*, “An investigation of noise shaping with perceptual weighting for WaveNet-based speech generation,” in *Proc.ICASSP*, 2018, pp. 5664–5668.

[15] A. van den Oord *et al.*, “Parallel WaveNet: Fast High-Fidelity Speech Synthesis,” *arXiv:1711.10433*, 2017.