Vol.2019-MUS-123 No.34
Vol.2019-SLP-127 No.34
2019/6/22

# 周期・非周期信号を用いた DNNに基づくリアルタイム音声ボコーダ

大浦 圭一郎[1,2,a)]　中村 和寛[2]　橋本 佳[1,2]　南角 吉彦[1]　徳田 恵一[1,2]

**概要**：本稿では，ニューラルネットワークに基づく音声ボコーダにおいて，周期信号と非周期信号を入力とする音声生成の枠組みを提案する．近年，ニューラルネットワークを用いて音声波形を直接モデル化する手法として WaveNet [1] が提案された．WaveNet は音声波形を高精度にモデル化することができ，自然な音声を直接生成することができるため，特に音声ボコーダ [2] として様々な研究で利用されている [3], [4], [5]．しかし，過去の音声サンプル列から次の音声サンプルを生成する自己回帰構造を持ち，合成時に並列演算ができないことから，実時間で合成できない問題があった．また，WaveNet を学習する際のデータベースに無い音高の再現ができない問題や，補助特徴量として指定したピッチ情報の音高を再現しないことがある問題があった．これらの問題に対し，本稿では明示的に周期信号と非周期信号の列を入力として用い，対応する音声サンプルの列を一度に生成する手法を提案する．提案手法を用いることで，実時間より高速に音声を生成できること，および，学習データの範囲外のピッチを持つ音声波形を生成できることを確認した．また，自然性に関する主観評価実験を行い，WaveNet と比較して合成音声品質の向上を確認した．

**キーワード**：DNN，敵対的学習，信号処理，音声合成，歌声合成

## 1. はじめに

音声は人間にとって最も親しみのあるメディアのひとつであることから，長らく様々な研究がなされてきた．近年では音声符号化，音声認識，音声合成などのデジタル信号処理に基づく音声関連技術がスマートフォンや家電等に導入され，人々の暮らしの中で利用されている．ここでいうデジタル信号処理とは，音声のアナログ信号から変換された離散時間信号を線形時不変システムの出力と仮定することにより，フーリエ変換や $z$ 変換等の理論に基づいて処理を行うものである．このような信号処理は，音源，声門，声道，放射モデル等によって構成される音声の生成モデルに基づいており，音声関連の研究分野では最も根本的な考え方として広く普及している．しかし，これまでの音声関連研究はこのような変換・処理で取り扱える枠組みの中に制限されていたため，モデル構造に関する過度の制約が性能の限界に繋がっていた．

このような流れの中で 2016 年に，自己回帰構造を持つ波形生成モデルである WaveNet [1] が提案された（図 1）．

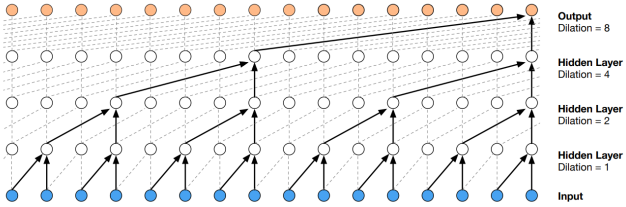


**図 1**　WaveNet

これは，モデル自身が過去に出力した数千の音声サンプル列から次の音声サンプルを予測するモデルであり，非常に高品質な音声を生成することができる．メルケプストラムや対数 F0 などの音響特徴量を補助特徴量として用いることで WaveNet を音声ボコーダ [2] として利用でき，近年の音声合成関連研究でしばしば利用されている [3], [4], [5]．WaveNet は自己回帰構造を持つため合成時に並列演算ができないことから，実時間で音声を合成できない問題がある．この問題を解決するため，WaveNet を教師モデルとして用い，自己回帰構造を持たない生徒モデルを学習することで，WaveNet と同等の品質の音声を高速に生成するモデルである Parallel WaveNet [6] が提案された．生徒モデルは教師モデルと同じ構造を持つが，教師モデルと異なり，モデル自身が生成した過去の音声サンプル列ではなく，ノイズ系列を入力として音声サンプルを出力する．学習は，

[1] 国立大学法人名古屋工業大学
Nagoya Institute of Technology, Nagoya, 466-8555, Japan
[2] 株式会社テクノスピーチ
Techno-Speech, Inc., Nagoya, 464-0858, Japan
[a)] uratec@nitech.ac.jp

ⓒ 2019 Information Processing Society of Japan

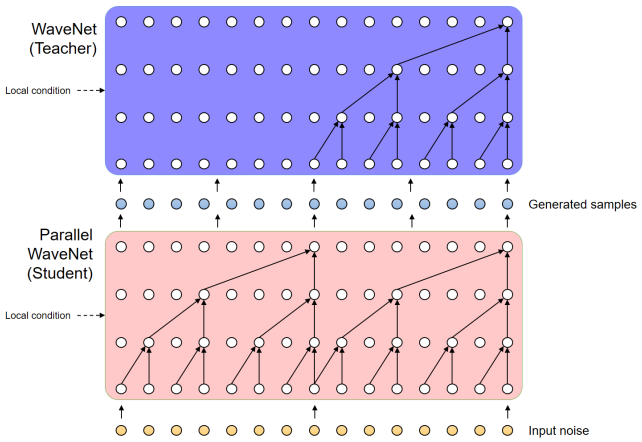


**図 2** Probability density distillation of Parallel WaveNet

自己回帰構造を持つ教師モデルの出力確率分布に順伝播型の生徒モデルの出力確率分布を近づける「蒸留」によって行われる（図 2）．Parallel WaveNet は自己回帰構造を持たないため，複数の音声サンプルを一度に出力することができ，WaveNet と比較して合成時間を大幅に削減することが可能になった．

自己回帰構造を持つ手法 [1], [7] や flow を用いた手法 [6], [8] は非常に品質の高い音声を合成することができるが，いくつかの問題がある．まず，これらのモデルは明示的な周期信号を入力に持たないことから，学習データの範囲外のピッチを持つ波形を生成できないだけでなく，たとえ学習データの範囲内のピッチを指定した場合でも異なるピッチの波形が生成されることがあった．これは，テキスト音声合成におけるイントネーションやピッチアクセントの品質劣化を引き起こす．特に歌声合成においては，ピッチの再現精度が品質に強い影響を及ぼすことから，解決すべき問題の一つである．また，Parallel WaveNet は WaveNet の合成時の高速化を実現したが，Probability density distillation loss 以外にも，複数の loss (power loss, perceptual loss, contrastive loss) を用いる必要があり，学習の際のチューニングが難しいという問題があった．

これらの問題に対し，我々はニューラルネットワークに基づく音声ボコーダにおいて，周期信号と非周期信号を入力とする音声生成の枠組み [9] を提案した．この手法では生成された音声波形の評価に $\mu$-law の離散確率分布を用いるが，他にも FFT スペクトルを用いる手法 [10], [11] などが提案されている．本稿では，音声が周期信号と非周期信号の和で成り立つと仮定し，より高品質な音声の生成手法を提案する．

## 2. 提案モデル

### 2.1 概要

提案モデルの概要を図 3 に示す．この構造は，近年活発に研究されている Deep Auto-Encoder (DAE) などの構造

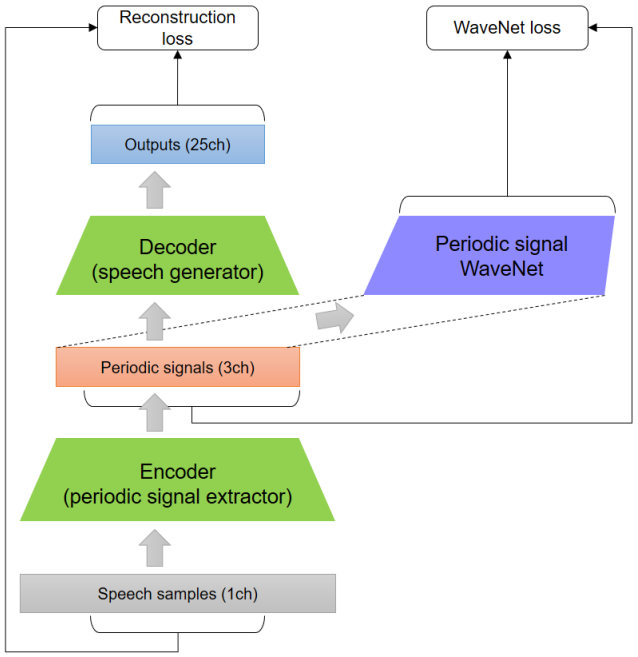


**図 3** Overview of the proposed model

と似たものになっており，音声波形を中間パラメータに変換する Encoder (periodic signal extractor) と中間パラメータを音声に変換する Decoder (speech generator) を直列に接続した構造である．中間パラメータを表現する WaveNet をあらかじめ学習しておき，Encoder (periodic signal extractor) の出力を評価する “WaveNet loss” と，入力された音声波形が適切に再現されることを評価する “Reconstruction loss” の合計を最小化するように Encoder/Decoder を同時学習する．合成の際は Decoder (speech generator) に任意の中間パラメータを入力することで，中間パラメータの周期性を保持した出力系列が得られることを期待する．

中間パラメータとしては，周期信号を表現する Sine 波および Cosine 波と有声無声情報の計 3 チャンネルを用いた．また，音声波形が周期信号と非周期信号の和から成り立つと仮定し，Decoder (speech generator) の出力は音声の周期信号 1 チャンネルと音声の非周期信号の帯域毎の強さ 24 チャンネルの計 25 チャンネルを用いた．音声波形の再現方法を図 4 右側に示す．まず Decoder (speech generator) から出力された 25 チャンネルの信号から，音声の非周期信号の帯域毎の強さ 24 チャンネルを分離し，帯域毎に分割したガウスノイズにかけ合わせることで音声の非周期信号を得る．これを音声波形の周期信号 1 チャンネルと足し合わせることで，周期・非周期が混合された音声波形が生成可能となる．

### 2.2 Reconstruction loss

音声波形が適切に再現されることを評価する “Reconstruction loss” としては，以下の 2 つを用いた．

**Gauss loss** Decoder (speech generator) の出力 25 チャ

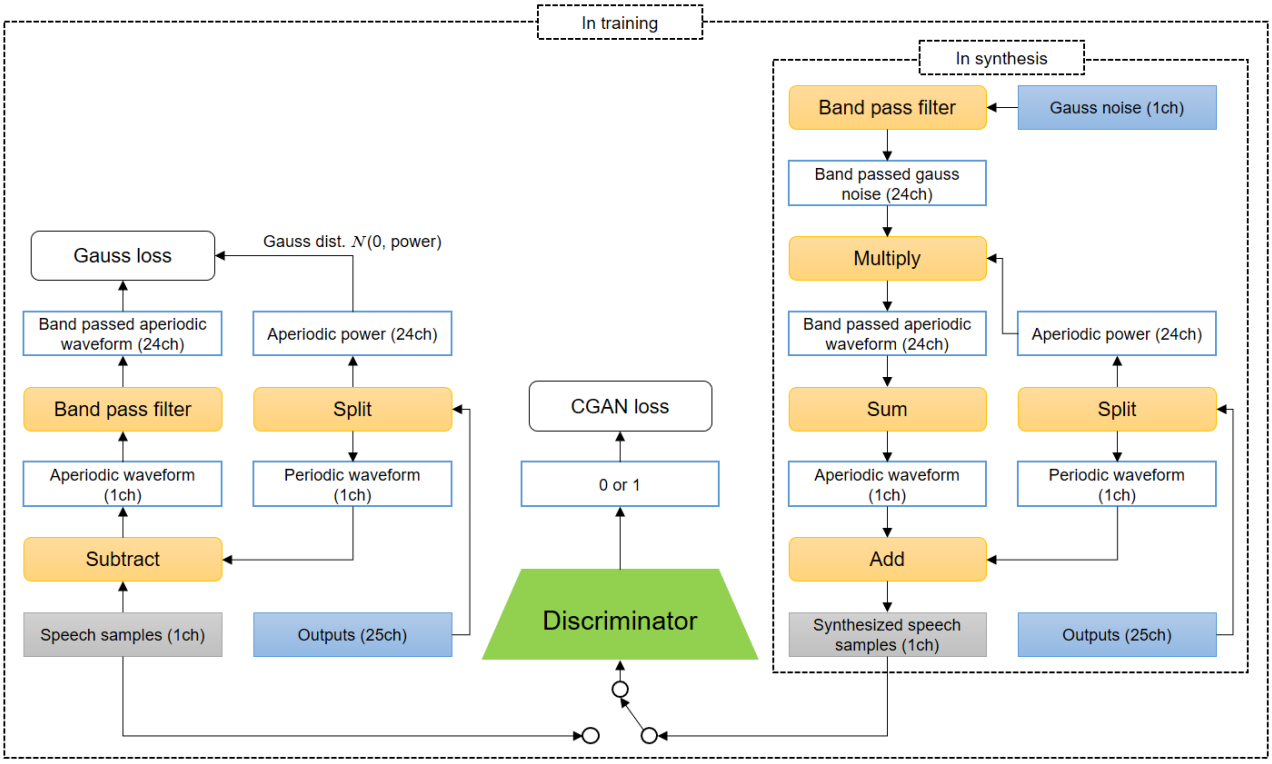


**図 4** Reconstruction loss calculation

ンネルから分離した音声の周期信号 1 チャンネルと自然音声の差を取ることで得られた非周期信号を帯域分割し，残りの 24 チャンネルを標準偏差とした 24 個のガウス分布で評価する（図 4 左側）．

**CGAN loss** 敵対的生成ネットワーク (Generative adversarial network; GAN) [12], [13] の枠組みを導入することで，自然音声と合成音声を識別する Discriminator を騙すロスを Encoder/Decoder の更新時に用いる．ただし，Discriminator の更新と Encoder/Decoder の更新は交互に行う．

### 2.3 モデル構造

WaveNet と提案モデルの詳細な構造をそれぞれ図 5 と図 6 に示す．WaveNet および Parallel WaveNet における入力はモデル自身が過去に出力した音声サンプル列や，ノイズ列の 1 チャンネルであるが，提案モデルでは明示的に周期信号（Sine 波と Cosine 波および有声無声情報の 3 チャンネル）を入力する．提案モデルは自己回帰構造を持たないことから，入力を過去だけの情報に制限する必要がないため，図 6 の通り，全ての層の Dilated convolution layer において，フィルタが左右対称になっており，現在の音声サンプルを出力するために，過去だけではなく未来の入力を用いる構造とした．Encoder/Decoder は最上層の Dilated convolution layer の出力を，全ての層から集約した skip layer の情報と結合している．Parallel WaveNet ではノイズから周期信号を生成する必要があるため，WaveNet の Dilated convolution layer より多くの Inverse-autoregressive flow (IAF) layer が必要になるが，提案モデルでは明示的に周期信号を入力するため，Dilated convolution layer の層数を WaveNet と同じとした．合成時，WaveNet では出力した離散確率分布に基づいてサンプリングすることで $\mu$-law のクラスを選択するが，提案モデルでは出力された 25 チャンネルの連続値をそのまま利用することで音声波形を生成する（図 4 の右側）．

Reconstruction loss に CGAN loss を用いる際の Discriminator の構造を図 7 に示す．ここでは，音声波形を入力とする 7 層の Dilated convolution layer の出力と補助特徴量から自然音声と合成音声を識別する形にした．

## 3. 実験

提案法の有効性を示すため，70 曲の日本語童謡歌声データベースを用いてモデル化性能を比較した．音声波形のサンプリング周波数は 48kHz，量子化ビット数は 16bit である．学習には 60 曲（約 65 分）を用い，テストに 10 曲（約 6.4 分）を用いた．WaveNet および Encoder/Decoder の構造は図 5 と図 6 に示した通りである．ただし，提案法において Decoder (peroidic signal extractor) の出力を評価する WaveNet の出力は，従来の WaveNet における 256 チャンネルの離散確率分布ではなく，中間パラメータを表す 3 チャンネルとした．WaveNet，Encoder/Decoder および Discriminator の補助特徴量としては WORLD [14] で分析した 50 次元のメルケプストラム係数，50 次元の非周期

ⓒ 2019 Information Processing Society of Japan

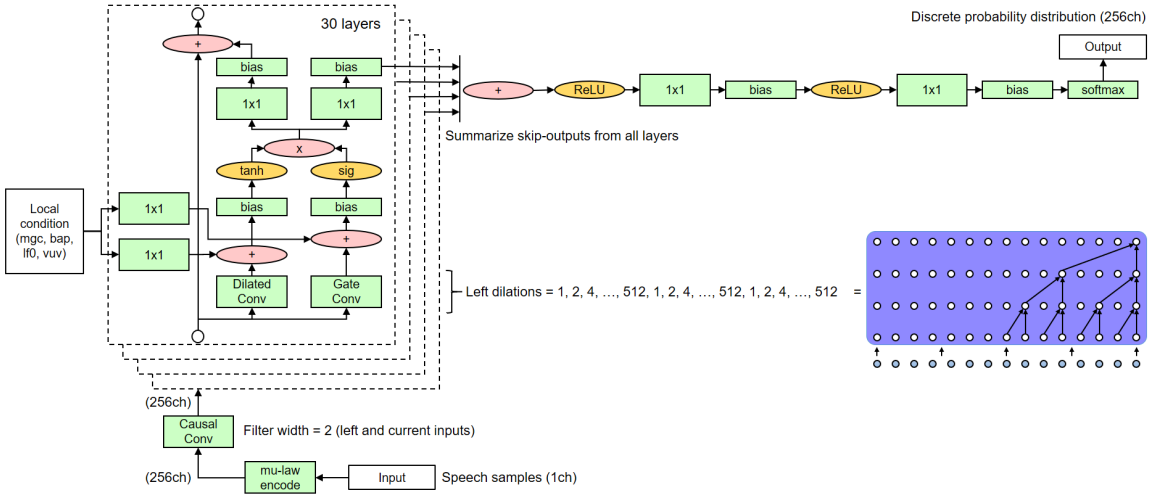


**図 5** WaveNet structure

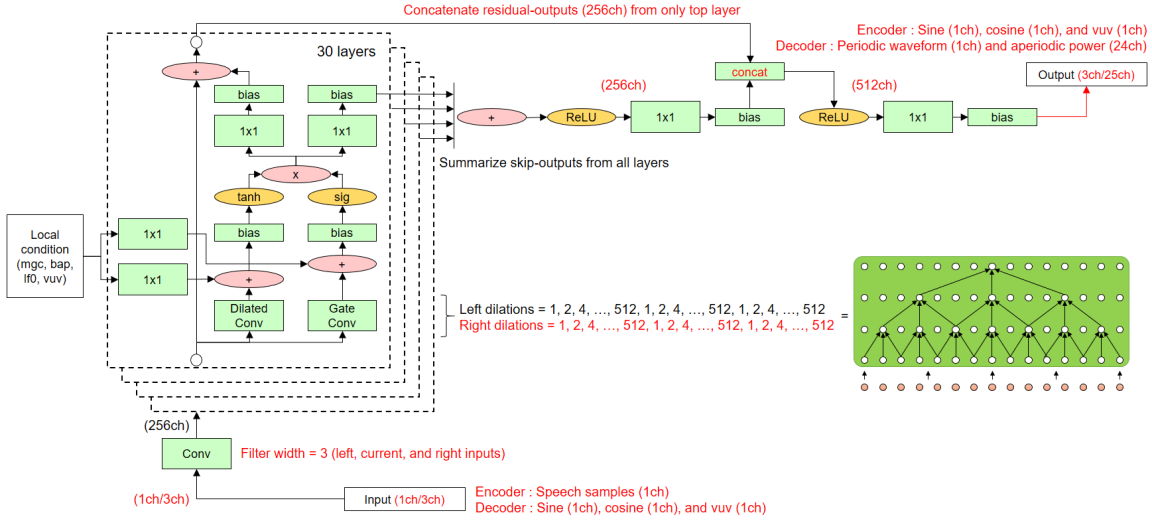


**図 6** Encoder/decoder structure of the proposed framework

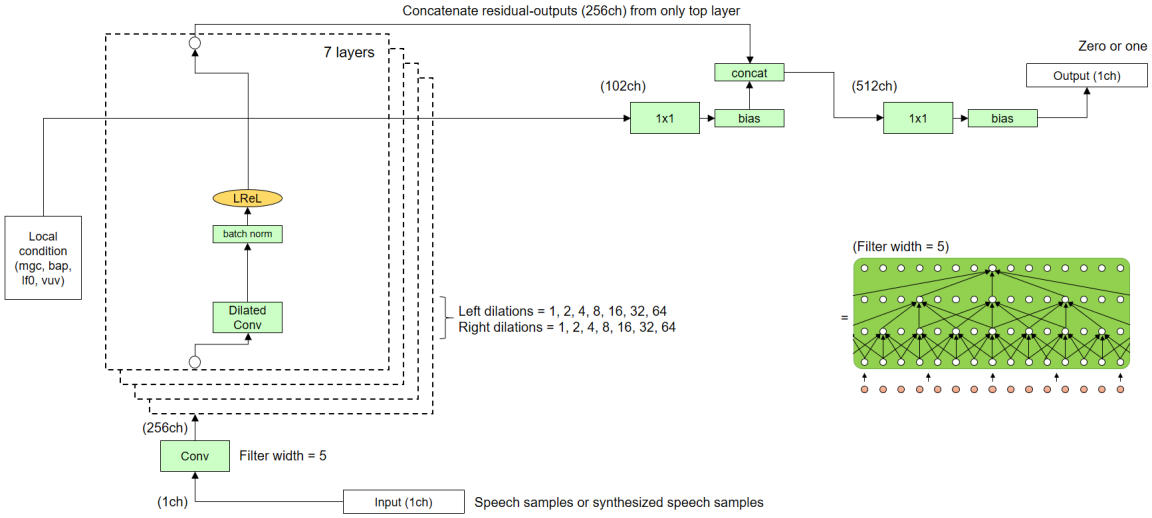


**図 7** Discriminator structure of the proposed framework

ⓒ 2019 Information Processing Society of Japan

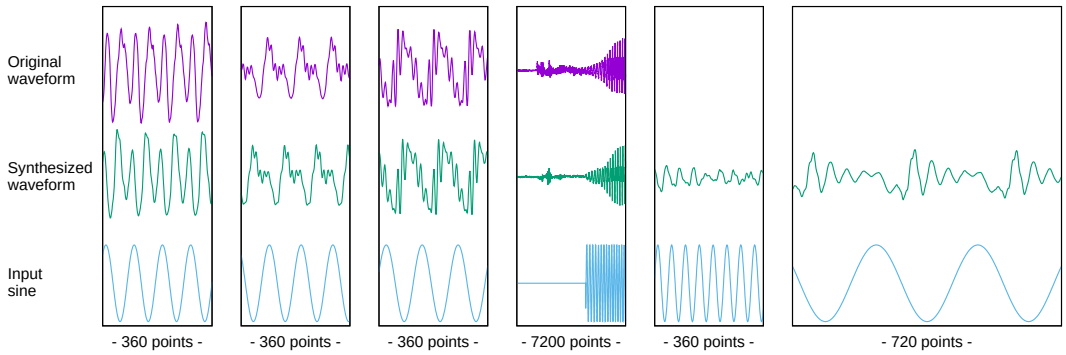


**図 8** Comparison with original waveform (upper), synthesized waveform (middle), and input sine (lower). /a/, /o/, /i/, and /ko/ in the test data, /a/ and /e/ exceeded over/under the pitch range are shown from the left.

性指標，対数基本周波数，有声無声情報の計 102 次元を用いた．

本実験では学習の補助のため，Encoder/Decoder の初期学習を行った．声門閉鎖点抽出ソフトウェアである REAPER [15] によって抽出した声門閉鎖点を基準として Sine 波および Cosine 波を生成し，音声波形に対応した周期信号として用いることで，Encoder (speech generator) と Decoder (periodic signal extractor) を個別に初期学習した．学習アルゴリズムには学習率を 0.0001 とした Adam を用いた．また，提案手法における合成時間を計測したところ，NVIDIA GTX 1080 を用いて実時間の約 5 分の 1 の速度で生成できることを確認した．

まず，提案法によって生成された波形を図 8 に示す．上段が自然音声，中段が合成波形，下段が入力した周期信号の Sine 波である．左側 4 つの図より，テストデータにおいて，入力した周期信号と同じピッチを持ち，かつ，自然波形と似た音声波形が生成できていることがわかる．右側 2 つの図は，テストデータをそれぞれ 1 オクターブ上げ下げすることで，学習データの範囲外の周期信号を入力した結果である．合成音声の自然性は低下したが，ピッチは保持されていることを確認した．

次に，16 人の被験者に対して，各手法につき 10 フレーズをテストデータからランダムに聞かせ，自然性を対象とした 5 段階評価を実施した．提案法としては，Reconstruction loss として Gauss loss だけを用いた手法，CGAN loss を加えた手法の 2 つを用い，比較手法として WaveNet および自然波形を用いた．

実験結果を図 9 に示す．図より，提案法は WaveNet より高い自然性を達成したことが確認できる．一方，提案法において，CGAN loss を用いることによる自然性の向上は確認できなかった．これは，合成音声が 32bit の単精度浮動小数点実数であるのに対し，自然音声として 16bit の符号付整数に量子化された音声波形を 32bit の単精度浮動小数点実数に変換して Discriminator に入力したことから，

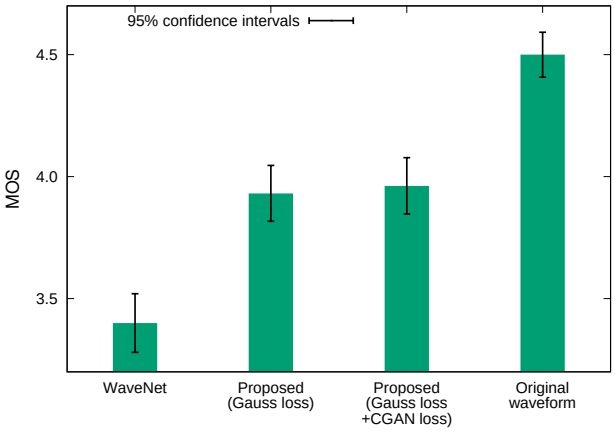


**図 9** Mean opinion scores for the naturalness

Discriminator が自然音声と合成音声の識別ではなく量子化の有無を識別している可能性が考えられる．

## 4. おわりに

本稿では，ニューラルネットワークに基づく音声ボコーダにおいて，周期信号と非周期信号を入力とする音声生成の枠組みを提案した．提案法では，周期信号と非周期信号の列を入力することで，対応する音声サンプルの列を一度に生成する．提案手法を用いることで，実時間より高速に音声波形を生成できること，および学習データの範囲外のピッチを持つ音声波形を生成できることを確認した．また，主観評価実験により，WaveNet と比較して音声品質の向上を確認した．今後の課題として，より大きなデータベースでの実験，不特定話者実験などが挙げられる．

**謝辞**

カシオ科学振興財団，JSPS 科研費 JP18K11163

## 参考文献

[1] A. van den Oord, *et. al.* “WaveNet: A generative model for raw audio,” *CoRR*, vol. abs/1609.03499, 2016.
[2] A. Tamamori, *et. al.* “Speaker-dependent WaveNet vocoder,” *INTERSPEECH* 2017, pp. 1118–1122, 2017.

ⓒ 2019 Information Processing Society of Japan

[3] J. Shen, *et. al.* “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” *CoRR*, vol. abs/1712.05884, 2017.
[4] J. Niwa, *et. al.* “Statistical voice conversion based on WaveNet,” *ICASSP* 2018, pp. 5289–5293, IEEE, 2018.
[5] K. Sawada, *et. al.* “The nitech text-to-speech system for the blizzard challenge 2018,” *Blizzard Challenge 2018 Workshop*, 2018.
[6] A. van den Oord, *et. al.* “Parallel WaveNet: Fast high-fidelity speech synthesis,” *CoRR*, vol. abs/1711.10433, 2017.
[7] N. Kalchbrenner, *et. al.* “Efficient Neural Audio Synthesis,” *CoRR*, vol. abs/1802.08435, 2018.
[8] R. Prenger, *et. al.* “WaveGlow: A Flow-based Generative Network for Speech Synthesis,” *CoRR*, vol. abs/1811.00002, 2018.
[9] 大浦, 他, “周期・非周期信号から駆動するディープニューラルネットワークに基づく音声ボコーダ,” 日本音響学会春季研究講演論文集, pp. 1049–1052, 2019.
[10] X. Wang, *et. al.* “Neural source-filter-based waveform model for statistical parametric speech synthesis,” *ICASSP* 2019, pp. 5916–5920, IEEE, 2019.
[11] O. Watts, *et. al.* “Speech waveform reconstruction using convolutional neural networks with noise and periodic inputs,” *ICASSP* 2019, pp. 7045–7049, IEEE, 2019.
[12] I. Goodfellow, *et. al.* “Generative adversarial nets,” in *NIPS*, 2014.
[13] M. Mirza, *et. al.* “Conditional Generative Adversarial Nets,” CoRR, vol. abs/1411.1784, 2014.
[14] M. Morise, *et. al.* “WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications,” *IEICE Transactions on Information and Systems*, vol. E99. D, no. 7, pp. 1877–1884, 2016.
[15] “REAPER: Robust epoch and pitch estimator,” https://github.com/google/REAPER.

ⓒ 2019 Information Processing Society of Japan