リアルタイムニューラルボコーダにおける学習データ量の影響の調査*

☆ 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]，志賀芳則[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

近年，ニューラルネットワークによる音声合成技術の発展により，テキスト音声合成は人間の声と区別がつかないほど高品質な音声を合成できるようになってきている [1]。これらの発展の大きな転換点として，直接波形生成モデル WaveNet [2] の登場が挙げられる。WaveNet の登場以前は，テキストから音響特徴量を推定する音響モデルにはニューラルネットワークが用いられていたが，音響特徴量から波形に変換するボコーダに関してはソースフィルタに基づく手法が用いられており，ボコーダの処理の部分で肉声感が失われてしまうという問題があった。そこで WaveNet ボコーダ [3] をはじめとするニューラルボコーダを用いることによりそれらの問題を解決し，従来手法の品質を大きく上回るテキスト音声合成を実現した。

WaveNet には合成速度が遅いため応用先が制限されるという問題点があったが，WaveNet より高速な新しいニューラルボコーダモデルが数多く提案されてきており，CPU でもリアルタイム合成が可能なモデルも登場してきている [4, 5]。これらの先行研究では数時間から数十時間のデータを用いて学習されているが，実際の課題解決に応用する場合は大量の学習データを用意できないことが考えられる。データ量の少ない話者の例として，障害により発話が不安定になってしまっている構音障害者が挙げられる。構音障害者は発話動作による本人の負担が大きいため，大量の音声データを用意することができない。構音障害者の発話を支援することは障害による社会活動の格差を低減し，よりよい社会形成の一助になると考えられる [6]。ニューラルボコーダに必要な学習データ量は WaveNet ボコーダでの調査がされている程度であり [7]，リアルタイムニューラルボコーダに必要なデータ量の調査は行われていない。

本稿では，少量の学習データしかないような場合でもリアルタイムニューラルボコーダにより高品質な音声を合成できるかを検証するため，WaveNet，LPCNet [4]，Parallel WaveGAN[5] を 1/8 時間から 9 時間のデータで学習させ，品質の変化に対する主観評価を行う。その際，より高品質な合成を実現するため [1, 8]，従来はサンプリング周波数 16 kHz であった LPCNet を 24 kHz へと拡張させる。また，LPCNet と Parallel WaveGAN における CPU コア数による合成速度の変化についても考察する。

2 ニューラルボコーダ

本稿では WaveNet，LPCNet，Parallel WaveGAN の 3 つのニューラルボコーダを採用した。WavaNet はリアルタイム生成はできないが，後者の 2 つのモデルはリアルタイム合成が可能である。

2.1 WaveNet

WaveNet は自己回帰機構を用いて波形を直接推定するニューラル生成モデルである。一般的な音声信号は 16 bit 整数値であるが，WaveNet では予測を簡単にするために μ-law 変換を適用し 8 bit 整数値に圧縮している。また波形予測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，自己回帰モデルであることと，数十層の畳み込みニューラルネットワークを用いた巨大なネットワークであることから生成速度が遅く，応用先が制限されてしまうという問題点がある。

2.2 LPCNet

LPCNet は再帰的ニューラルネットワークを用いた WaveRNN[9] ベースのニューラルボコーダであり，線形予測分析 (linear prediction coding: LPC) した残差信号をニューラルネットワークで推定する。WaveRNN が Dual Softmax により 16bit 波形を直接生成するのに対して，LPCNet は WaveNet と同様に μ-law 圧縮した 8bit 信号を生成するが，残差信号であるため量子化誤差が少なく，結果として WaveRNN よりも少ないモデルパラメータで高品質な音声合成が可能である [4]。Fig. 1 に概略図を示す。LPCNet は，入力の音響特徴量から特徴量抽出を行う Frame rate network と，LPC で予測された音声波形サンプルから残差信号を推定する Sample rate network の 2 つのモデルから構成される。ここで，サンプリング周波数が 16 kHz の場合，入力特徴量は 18 次元のバーク尺度のケプストラムとピッチ周期，ピッチ相関である。また，LPC

*Investigation of the effect of amount of training data for real-time neural vocoders by MATSUBARA, Keisuke[1,2], OKAMOTO, Takuma[2], TAKASHIMA, Ryoichi[1], TAKIGUCHI, Tetsuya[1], TODA, Tomoki[3,2], SHIGA, Yoshinori[2] and KAWAI, Hisashi[2] ([1]Kobe Univ, [2]NICT, [3]Nagoya Univ)

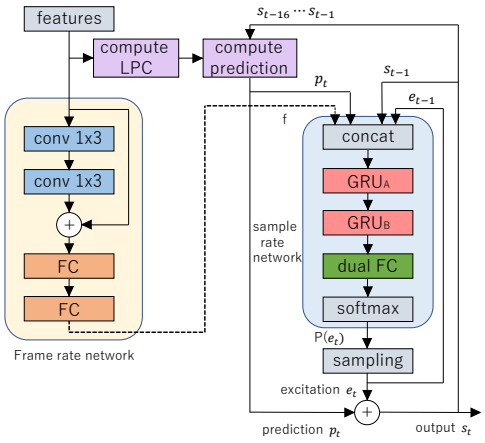

Fig. 1　LPCNet structure.

による予測波形は以下の式で表される。

$$p_t = \sum_{k=1}^{M} a_k s_{t-k} \tag{1}$$

ここで p_t は t ステップ目の予測波形サンプル，a_k は k 次 LPC 係数，s_{t-k} は $t-k$ ステップ目の実際の波形である。LPC 係数は入力特徴量のバーク尺度のケプストラムから計算される。具体的には，まずバークケプストラムから通常の周波数尺度のパワースペクトル密度を計算する。次にパワースペクトル密度を逆フーリエ変換することで自己相関関数を計算する。最後に自己相関関数からレビンソンダービン法を用いて LPC 係数を計算する。LPC の予測波形サンプルと，音響特徴量を入力した Frame rate network の出力，1 ステップ前の実際の音声波形サンプルと予測残差波形サンプルを入力として，現在の予測残差波形サンプルを Sample rate network で予測する。Sample rate network は Gated Recurrent Unit(GRU) を用いた自己回帰モデルであるため，本来は WaveNet と同様に生成に時間を要するが，Sparse coding と呼ばれるネットワークの重み行列のうち値が小さいものをゼロに置き換える手法を用いることで高速化を行っており，モバイル CPU でもリアルタイム生成が可能となっている。

2.2.1　サンプリング周波数 24 kHz への拡張

LPCNet は従来サンプリング周波数 16 kHz の音声を合成可能なニューラルボコーダとして提案されているが，更に高品質な 24 kHz の音声を合成するために入力特徴量の修正を行う。従来手法では 18 次元のバークケプストラムとピッチ周期，ピッチ相関の計 20 次元を入力特徴量としていたが，ここではバークケプストラムの次元数を 20 に拡張する。

バーク尺度でのフィルタバンク分割は，音声圧縮手法の Opus[10] に準拠して行う。Fig. 2 にフィルタバンク構造を示す。この手法では低周波領域ではバーク尺度の大きさによらず一定の間隔で分割していき，高周波領域ではバーク領域での間隔が一定となるように分割している。16 kHz でサンプリング可能な 8 kHz までをカバーするには低次 18 次元，24 kHz でサンプリング可能な 12 kHz までカバーするには低次 20 次元までを用いる。

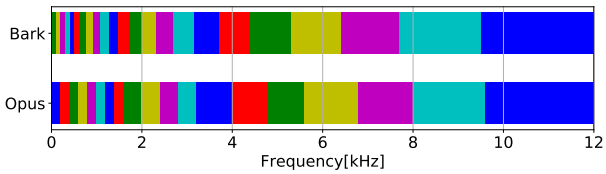

Fig. 2　Opus and Bark band layout.

2.3　Parallel WaveGAN

Parallel WaveGAN は，Parallel WaveNet を Generator とする敵対的生成ネットワーク (Generative Adversarial Network: GAN) をベースとするニューラルボコーダである。入力はホワイトノイズと音響特徴量であり，Generator である WaveNet が全て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を偽と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) を行った場合の誤差である 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であり，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問題点があるが，ここでは解像度の異なる複数の STFT で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 は自己回帰機構を持たないため，一度に複数のサンプルを同時に生成することで高速生成を可能にしており，CPU でもリアルタイム生成が可能である[1]。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。

3　実験

3.1　実験条件

WaveNet，LPCNet，Parallel WaveGAN ボコーダにおける学習データ量と品質の関係を調査するため，サンプリング周波数 24 kHz の音声を用いた分析合成を行う。データには JSUT コーパス [11] より日本人女性話者による 7697 文 (約 10 時間) の音声を使用した。また大語彙連続音声認識エンジンである Julius[12] を用いて音素アライメントを行い，音声データの前後に

[1] https://github.com/kan-bayashi/ParallelWaveGAN

含まれる無音区間の除去を行った。テストセットと検証セットに 100 文ずつ，学習セットには最大 7497 文（約 9 時間）を使用した。学習データ量の種類としては，LPCNet，Parallel WaveGAN は 9，5，3，1，1/2，1/4，1/8 時間の 7 種類，WaveNet は 9 時間と 1/8 時間の 2 種類を評価した。また比較として，WORLD ボコーダ [13](D4C edition [14]) と 9 時間のデータ量で学習させた WaveGlow[15] ボコーダも評価に加えた。

WaveNet ボコーダは [3] と同様のモデルを使用した。また予測誤差による雑音成分を抑えるため，時不変ノイズシェーピング法 [16] を適用した。入力特徴量にはメルスペクトログラム 80 次元を使用した。メルスペクトログラムの計算では 42.7 ms の Hann 窓を用いて，フレームシフトを 5 ms とした。

LPCNet は [4] と同様のモデルを使用した。入力特徴量にはバーク尺度の 20 次元ケプストラムとピッチ周期，ピッチ相関を用いた。バークケプストラムの計算に際しては，20 ms の Vorbis 窓を用い，フレームシフトを 10 ms としてスペクトル分析を行い，バーク尺度を用いたフィルタバンクを適用した後に離散コサイン変換を行った。ピッチの計算にはオープンループの相互相関関数をベースとする手法を用いた。

Parallel WaveGAN は [5] と同様のモデルを使用した。入力特徴量にはメルスペクトログラム 80 次元を使用した。メルスペクトログラムの計算では 20 ms の Hann 窓を用いて，フレームシフトを 12.5 ms とした。また周波数帯域を 80 から 7600 Hz に制限した。WaveGlow は [18] と同様のモデルを使用した。

パラメータの更新は，WaveNet，LPCNet は Adam を使用し，Parallel WaveGAN と WaveGlow では RAdam[17] を用いた。学習には NVIDIA V100 の GPU を使用し，WaveNet，LPCNet，Parallel WaveGAN，WaveGlow の学習にそれぞれ 4 枚，8 枚，2 枚，4 枚使用した。

主観評価には，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 18 人の成人日本語母語話者で，テストセット 20 文に対して 18 条件と原音の計 19 条件の 380 文をヘッドホン聴取により評価した。

3.2 実験結果

Table 1 に WaveNet，LPCNet，Parallel WaveGAN，WaveGlow の生成速度を示す。また，Fig. 3 に LPCNet と Parallel WaveGAN における CPU コア数による生成速度の変化を示す。WaveNet は NVIDIA の TeslaV100 の GPU を，LPCNet と Parallel WaveGAN は Intel の Xeon6152 の CPU を用いて合成した。Table 1 の結果より，LPCNet はシングル CPU で RTF=0.24 の高速な生成が可能であることが分かった。また Parallel WaveGAN はシングル CPU で RTF=2.38 となり，リアルタイム生成に届かない結果となった。しかし Fig. 3 の結果より CPU コア数を増やすことで生成速度が改善され，16 コアで最大 RTF=0.41 まで高速化することができた。LPCNet は CPU コア数による生成速度の改善は見られなかった。これは，Parallel WaveGAN が複数の波形サンプルを同時に生成できるのに対し，LPCNet は自己回帰モデルであるため，CPU コア数を増やしても並列処理が出来ないからであると考えられる。

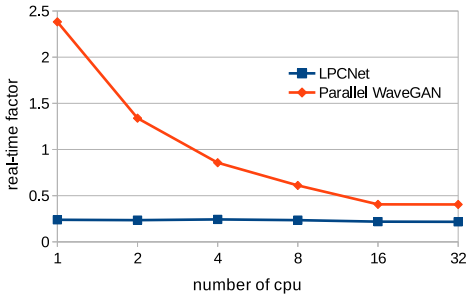

Fig. 3 Result of real-time factors for inference by increasing the number of CPU cores.

Table 1 Result of real-time factors for inference using an NVIDIA Tesla V100 or Intel Xeon 6152. (*) denotes number of CPU cores or GPUs.

model	RTF-CPU	RTF-GPU
WaveNet	-	196 (1)
LPCNet	0.24 (1)	0.22 (1)
Parallel WaveGAN	0.41 (16)	0.02 (1)
WaveGlow	-	0.07 (1)

聴取実験の結果を Fig. 4 に示す。学習データ量が 9 時間の場合は WaveNet が最も音質が良いという結果になった。学習データ量を減らしていくと，LPCNet，Parallel WaveGAN ともに音質が劣化していが，t 検定の結果，1 時間から 9 時間までは結果に有意な差は見られなかった。1/8 時間では，Parallel WaveGAN は WORLD の品質より悪くなったが，LPCNet は WORLD より良い品質を維持していることが分かった。結果として，LPCNet と Parallel WaveGAN は 1 時間程度の学習データ量で十分学習が可能と分かった。LPCNet が Parallel WaveGAN や WaveGlow より音質が良い理由としては，自己回帰モデルのため過去の波形情報も使えることや，予測残差は白色化しているため，結果として時不変ノイズシェーピング法 [16] と同様に予測誤差によるノイズが聴感上で抑制されていることなどが考えられる。また，今回の実験では WaveGlow の方が Parallel WaveGAN より若干音質が高かったが (t 検定により有意差あり)，

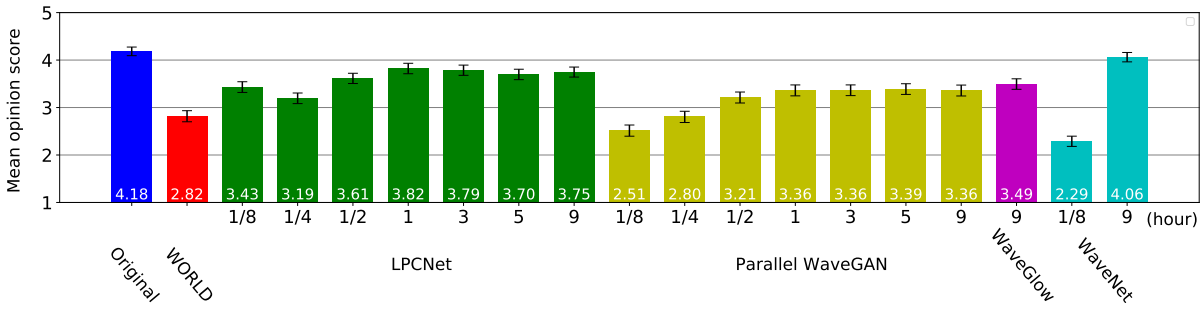

Fig. 4　Result of MOS test with 18 listening subjects. Confidence level of the error bars is 95 %.

Parallel WaveGAN は adversarial loss のハイパーパラメータを調整することにより品質が向上する可能性があり，これは今後の課題とする。

4　まとめ

WaveNet，LPCNet，Parallel WaveGAN ボコーダを少量の学習データで学習させ品質を評価した。結果として，LPCNet と Parallel WaveGAN は 1 時間程度の学習データで十分に学習ができることが示された。また生成速度の実験結果から，LPCNet と Parallel WaveGAN はシングル CPU あるいはマルチ CPU でリアルタイム生成が可能であることが示された。

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. van den Oord *et al.*, “WaveNet: A generative model for raw audio,” in *Proc. SSW9*, Sept. 2016, p. 125.

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

[4] J. Valin, J Skoglund, “LPCNet: Improving Neural Speech Synthesis Through Linear Prediction,” in *Proc. ICASSP*, May. 2019, pp. 5891–5895.

[5] R. Yamamoto *et al.*, “Parallel WaveGAN:A fast Waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” *arXiv:1910.11480*, 2019.

[6] 南阪ら，” 構音障害者の少量データを用いた深層学習による音声合成の検討”，音講論, pp. 1011–1014, Sept. 2019.

[7] 林ら，”WaveNet ボコーダにおける学習データ量の影響に関する調査”，音講論, pp. 249–250, Mar. 2018.

[8] A. van den Oord *et al.*, “Parallel WaveNet: Fast high-fidelity speech synthesis,” in *Proc. ICML*, July 2018, pp. 3915–3923.

[9] N. Kalchbrenner *et al.*, “Efficient Neutal Audio Synthesis,” in *Proc. ICML*, July 2018, pp. 2415–2424.

[10] J. Valin *et al.*, “High-quality, low-delay music coding in the Opus codec,” in *Proc. 135th AES Convention*, Oct. 2013.

[11] S. Takamichi *et al.*, “JSUT and JVS: free Japanese voice corpora for accelerating speech synthesis research,” *Acoust. Sci. Tech.*, (in press).

[12] A. Lee, T. Kawahara. “Recent Development of Open-Source Speech Recognition Engine Julius,” in *Proc. APSIPA ASC* Oct. 2009, pp. 131–137.

[13] 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.

[14] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” Speech Communication, vol. 84, pp. 57–65, Nov. 2016.

[15] R. Prenger *et al.*, “WaveGlow: A flow-based generative network for speech synthesis,” in *Proc. ICASSP*, May. 2019, pp. 3617–3621.

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

[17] L. Liu *et al.*, “On the Variance of the Adaptive Learning Rate and Beyond,” *Proc. ICLR*, Apr. 2020.

[18] T. Okamoto *et al.*, “Tacotron-based acoustic model using phoneme alignment for practical neural text-to-speech systems,” in *Proc. ASRU*, Dec. 2019, pp. 214–221.