Full-band LPCNet：48kHz リアルタイムニューラルボコーダ *

☆ 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

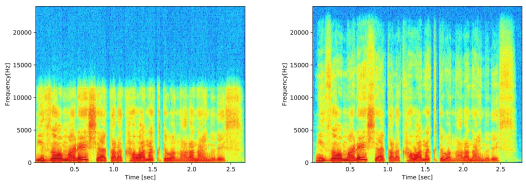
1 はじめに

テキスト音声合成 (Text-to-Speech: TTS) は，音声コミュニケーションの重要な技術の一つであり，盛んに研究が行われている。また近年では，ニューラルネットワークを用いた音声合成技術が急速に発展しており，自然性の高い音声を生成できるようになっている [1]。ニューラルネットワークによる TTS システムは，テキストを入力し，対応する音響特徴量を出力する音響モデルと，音響特徴量から対応する音声を出力するニューラルボコーダの 2 つのネットワークで構成される。特に，WaveNet ボコーダ [2] をはじめとするニューラルボコーダの登場は従来用いられていたソースフィルタボコーダ [3] による品質を大きく上回り，ニューラル音声合成技術の発展に大きく貢献している。

しかし WaveNet ボコーダは，巨大な畳み込み層で構成された自己回帰モデルであるため，合成速度が非常に遅いという問題を抱えていた。その問題を解決するために様々なニューラルボコーダモデルが提案され，高品質な音声をリアルタイムに合成できるようになってきている。中でも WaveRNN[4] や LPCNet[5] などは，CPU のみでリアルタイム合成が可能なニューラルボコーダとして提案されている。しかしこれらのモデルの多くは，最大でもサンプリング周波数 24 kHz の音声を生成することを想定している。Fig. 1 のように，人間の音声には 24 kHz 音声でサンプリング可能な 12 kHz よりも高周波の成分が含まれており，また人間の可聴領域は最大で 20 kHz 程度である。文献 [6] では，サンプリング周波数が 16 kHz の合成音声に対して，48 kHz 合成音声は主観評価による品質が改善されたと報告されている。

48 kHz 音声の生成を扱ったニューラル音声合成の研究はいくつかなされているが，それらの殆どはソースフィルタボコーダや WaveNet ボコーダを用いていた。前述の通り，ソースフィルタボコーダは音声の品質に限界があり，WaveNet ボコーダは生成速度の問題を抱えている。文献 [7] ではニューラルボコーダを用いたフルバンド音声のリアルタイム合成に成功しているが，学習手順が複雑であり，また合成には高性能な GPU が必要である。

そこで本稿では，サンプリング周波数 48 kHz のフルバンド LPCNet を提案する。これまでに, LPCNet の性能を改善する研究はいくつかなされているが，それらはサンプリング周波数を 16 kHz か 24 kHz としていた [8]。本稿ではサンプリング周波数を 48 kHz に拡張し，分析合成およびテキスト音声合成の両条件で品質の主観評価を行う。また CPU を用いた合成速度を測定し，48 kHz 音声合成のリアルタイム性を検証する。



(a) 24 kHz　　(b) 48 kHz

Fig. 1 Spectrograms of a female speech waveform with sampling frequencies of (a) 24 kHz and (b) 48 kHz.

2 LPCNet

LPCNet は WaveRNN に基づくニューラルボコーダであり，線形予測分析 (Linear Predinction Coding: LPC) で推定した音声信号の残差信号をニューラルネットで推定する。LPCNet は，入力された音響特徴量から特徴量抽出を行う Frame Rate Network と，LPC で推定された音声信号から残差信号を推定する Sample Rate Network の 2 つのブロックで構成されている。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 次元のバークケプストラムとピッチ周期，ピッチ相関である。

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。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。そこで，Sparse coding を用い

* Full-band LPCNet: A real-time neural vocoder for 48 kHz audio 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)

て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。

3 フルバンド LPCNet

3.1 入力特徴量の修正

16 kHz 合成の場合は 18 次元のバークケプストラムを用いていたが，48 kHz 合成のためにこれを 50 次元にまで拡張する。従来の LPCNet では，Opus[9] と呼ばれる音声圧縮手法に準じてフィルタバンク分割を行っている。Opus では，低周波領域では等間隔に，高周波領域ではバーク尺度に従うように周波数帯を分割する。最終的に，0 から 8 kHz までの周波数を 18 個のフィルタバンクに分割する。本稿では 48 kHz 音声の合成のため，24 kHz までの周波数をフィルタバンクに分割するが，Opus の手法より更に細分化したフィルタバンクを適用する。ここで，バーク尺度は以下の式で定義される。

$$B = 13\arctan(0.00076f) + 3.5\arctan((f/7500)^2) \tag{1}$$

ここで，f，B はそれぞれ線形周波数とバーク周波数である。本来のバーク尺度は，上述の式に従って 0 から 15.5 kHz の周波数帯を 24 のフィルタバンクへと分割する尺度として定義されており，よって 15.5 kHz 以上ではバーク尺度は本来定義されていないが，本稿では 48 kHz 音声でサンプリング可能な 24 kHz まで式 (1) が適用可能であると仮定し，0 から 24 kHz までの周波数帯を 50 個のフィルタバンクへと分割する。

3.2 モデルパラメータの修正

より高品質な 48 kHz 音声合成を実現するため，LPCNet のどのパラメータが品質に大きく影響するか調査した。モデルパラメータを増加させれば品質は向上するが，その分合成速度が増加しリアルタイム性が損なわれるため，調整は一部のパラメータを変更するのみにとどめて調査を行った。予備実験の結果，Sample Rate Network に用いられる $\mathrm{GRU_A}$ のチャネル数が品質に大きく影響することが分かった。そこで本稿では，$\mathrm{GRU_A}$ のチャネル数を 384(オリジナル)，512，640 の 3 種類を用いて実験を行った。

4 実験

4.1 実験条件

フルバンド LPCNet による合成音声の品質を評価するため，分析合成およびテキスト音声合成での主観評価を行う。比較対象には，WORLD ボコーダ [3]，WaveNet ボコーダ [2] および Parallel WaveGAN[10] を用いた。データセットは，JSUT コーパス [11] より，日本人女性話者 1 名による 7697 文 (約 10 時間) の 48 kHz 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含まれる無音区間の除去を行った。ニューラルボコーダの学習には 7,497 文を使用した。またニューラル音響モデルの学習には，人手によりラベリングされた基づく HTS 形式のコンテキストラベルが利用可能な 4,800 文を使用した。[1] テストセットと検証セットにはそれぞれ 100 文ずつを使用した。

LPCNet は [5] の実装を元に，$\mathrm{GRU_A}$ のチャネル数を 384，512，640 と変化させて実験を行った。入力特徴量には 50 次元のバーク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，窓長を 20 ms，フレームシフトを 10 ms としてスペクトル分析を行い，バーク尺度による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。

WaveNet は [2] 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 [13] を適用した。入力特徴量には周波数帯域を 0 から 7.6 kHz に制限した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，窓長を 42.7 ms，フレームシフトを 5 ms とした。

Parallel WaveGAN は [10] と同様のモデルを使用した。入力特徴量には WaveNet と同じく，帯域制限を施した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窓長を 20 ms，フレームシフトを 5 ms と 12.5 ms の 2 種類を用意して実験を行った。また本来は 24 kHz 音声合成のモデルとして提案されている Parallel WaveGAN を 48 kHz 音声合成に対応するため，内部で用いられている STFT 損失のフレーム長，窓長及びフレームシフトを 2 倍の大きさに調整した。

テキスト音声合成では，フルコンテキストラベルを入力とする Tacotron ベースの音響モデル (BLSTM+Tacodec) と Tacotron2 を用いた [14, 15]。入力特徴量は 575 次元のコンテキストラベルから，38 次元の音素と HTS ラベルにおける A と F から 9 次元のアクセントラベルを抽出した 47 次元のベクトルとした。BLSTM+Taco2dec で必要となる音素継続長についてもコンテキストラベルから抽出して使用した。出力特徴量は後段のボコーダによって異なり，フルバンド LPCNet，WORLD，WaveNet ボコーダ，

[1] https://github.com/sarulab-speech/jsut-label

Table 1 Real-time factors for inference using an NVIDIA Tesla V100 GPU or Intel core i7-8700B (single core) or Intel Xeon 6152 (multi cores) CPUs. (*) denotes the number of CPU cores.

model	GPU	CPU	Multi CPU
BLSTM+Taco2dec	0.08	0.42	0.18(8)
Tacotron2	0.08	0.37	0.18(8)
WaveNet	672	-	6144(16)
Parallel WaveGAN	0.03	-	1.1(16)
LPCNet[384]	0.46	0.42	-
LPCNet[512]	0.62	0.61	-
LPCNet[640]	0.95	0.84	-

Parallel WaveGAN でそれぞれ 52，57，80，80 次元の音響特徴量とした。パラメータの更新は，LPCNet では Adam を，それ以外のニューラルボコーダ及び音響モデルでは RAdam[16] を使用した。

4.2 実験結果

4.2.1 合成速度の比較

Table 1 に，NVIDIA Tesla V100 GPU，Intel core i7-8700B (単一 CPU コア)，Intel Xeon 6152(複数 CPU コア) を用いた，各モデルにおける合成速度をリアルタイムファクター (RTF) として示す。ここでは，LPCnet は C 言語での実装，それ以外の音響モデル及びニューラルボコーダは PyTorch による実装を用いている。Table 1 の結果より，LPCNet は 48 kHz 音声合成であっても，単一 CPU でリアルタイム合成を達成している。ここで WaveNet と Parallel WaveGAN の RTF は，CPU コア数を増やしながら計測した中で最も良かったものを提示している。Parallel WaveGAN は 24 kHz 音声合成では 16CPU コアで RTF=0.41 であったが [8]，48 kHz 音声の合成では RTF が 1 を下回ることはなかった。Parallel WaveGAN の合成速度をさらに改善するには，LPCNet と同様に C 言語による実装が必要と考えられる。LPCNet でも複数 CPU での RTF を計測したが，単一 CPU の場合と比較して明確に改善されることはなかった。音響モデルと LPCNet を合わせた TTS システム全体で考えた場合は，単一 CPU ではリアルタイム合成に届かなかった。しかし音響モデルは CPU コア数を増やすことで速度が改善され，最大 8 コアまで改善が見られた。結果として，8CPU コアを用いることでフルバンド LPCNet によるリアルタイムテキスト音声合成システム実現できた。本実験では，音響モデルの実装に PyTorch を用いていたため，LPCNet と同様に C 言語による実装を行うことで更に速度が改善できると考えられる。また近年では，自己回帰モデルを用いない高品質なニューラル音響モデルが提案されており [17]，それらを用いることでもさらなる改善ができると考えられる。

4.2.2 主観評価の結果

主観評価として，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 25 人の成人日本語母語話者で，テストセット 20 文に対して 20 条件の計 400 文をヘッドホン聴取により評価した。原音の他に 8bit の μ-law 圧縮を適用した音声と，LPC 分析した残差信号を 8bit-μ-law 圧縮した音声 (LPCNet のオラクル条件) の 3 条件を用いた。また Tacotron2 によるテキスト音声合成の実験は，フルバンド LPCNet に対してのみ行った。

Fig. 2 に聴取実験の結果を示す。分析合成では WaveNet ボコーダが最も高いスコアとなった。それに次いで，$\mathrm{GRU_A}$ チャネル数が 640 のフルバンド LPCNet が高いという結果になったが，WaveNet には合成速度の問題があるため，LPCNet の優位性は十分にあるといえる。また $\mathrm{GRU_A}$ チャネル数を減らしていくと，比例して品質が劣化していったが，384 チャネルの場合でも WORLD や Parallel WaveGAN の品質には勝るという結果となった。今回の実験では，Parallel WaveGAN が最も品質が悪いという結果になったが，これは元々24 kHz 音声合成が可能なモデルとして提案されたモデルをパラメータ調整等を行わずに実験したことが原因と考えられる。よってサンプリング周波数の増加に合わせて適切にパラメータ調整を行うことで品質を改善できると考えられ，これは今後の課題とする。

テキスト音声合成では，$\mathrm{GRU_A}$ チャネル数が 640 の LPCNet が分析合成の品質を上回り，原音の品質に匹敵する結果となった。WaveNet は分析合成での品質より大きく劣化する結果となった。これは音響モデルから出力される音響特徴量が過剰に平滑化されてしまっていることが原因として考えられる。LPCNet はテキスト音声合成での劣化が見られなかったが，これは LPCNet が入力として用いる音響特徴量が違うことが理由の 1 つとして挙げられる。WaveNet では入力特徴量はメルスペクトログラムであったが，LPCNet の入力特徴量は声道特徴に対応するバークケプストラムと，声帯振動に対応する基本周波数を用いている。そして LPCNet で用いたような音響特徴量の方が，音響モデルにとって推定がしやすく，48 kHz 音声のような高品質な音声合成にも耐えう劣化にとどまったと考えられる。また LPCNet は，残差信号を推定することで量子化誤差を抑制させている点や，学習時に入力特徴量にノイズを加える点など，頑健性を高めるための処理が施されており，これらの処理も

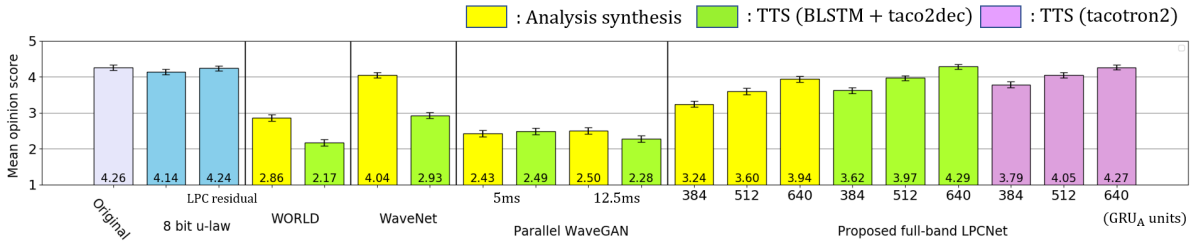


Fig. 2　Result of MOS test with 25 listening subjects. Confidence level of the error bars was 95 %.

テキスト音声合成による劣化を低減させていると考えられる。

5　まとめ

本稿では高品質な 48 kHz 音声を合成可能なフルバンド LPCNet を提案しその性能を評価した。実験の結果，複数 CPU コアを用いることでフルバンド LPCNet によるリアルタイムテキスト音声合成システムを実現した。またテキスト音声合成による合成音声の品質は原音の品質に匹敵する結果となった。

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. Tamamori *et al.*, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, Aug. 2017, pp. 1118–1122.

[3] 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.

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

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

[6] X. Wang *et al.*, "A Comparison of Recent Waveform Generation and Acoustic Modeling Methods for Neural-Network-Based Speech Synthesis," in *Proc. ICASSP*, Apr. 2018, pp. 4804–4808.

[7] K. Oura *et al.*, "Deep neural network based real-time speech vocoder with periodic and aperiodic inputs," in *Proc. SSW10*, Sept. 2019, pp. 13–18.

[8] 松原ら, "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", 音講論, Mar. 2020, pp. 1045–1048.

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

[10] R. Yamamoto *et al.*, "Parallel WaveGAN:A fast Waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, May 2020, pp. 6199–6203.

[11] S. Takamichi *et al.*, "JSUT and JVS: free Japanese voice corpora for accelerating speech synthesis research," in *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] 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.

[14] 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.

[15] T. Okamoto *et al.*, "Transformer-based text-to-speech with weighted forced attention," in *Proc. ICASSP*, May 2020, pp. 6729–6733.

[16] L. Liu *et al.*, "On the Variance of the Adaptive Learning Rate and Beyond," in *Proc. ICLR*, Apr. 2020.

[17] Z. Zeng *et al.*, "AlignTTS: Efficient feed-forward text-to-speech system without explicit alignment," in *Proc. ICASSP*, May 2020, pp. 6741–6718.