Vol.2019-SLP-129 No.3
2019/10/11

[チュートリアル講演] 音声波形直接生成モデル「ニューラルボコーダ」の比較

岡本 拓磨[1,a)]　戸田 智基[2,1]　志賀 芳則[1]　河井 恒[1]

概要：2016 年までの統計的テキスト音声合成や声質変換では，ニューラルネットワークに基づく音響モデルを用いたとしても，ソースフィルタモデルに基づくボコーダによる音質劣化が肉声感を阻む大きな壁となっていた．2016 年 9 月，WaveNet からはじまるニューラルボコーダの登場により，言語特徴量や音響特徴量からニューラルネットにより音声波形を直接合成できるようになり，Tacotron 2 においては，ついに自然音声と区別のつかない品質の英語テキスト音声合成が実現された．現在，ニューラルボコーダは音声合成における基盤技術となり，様々な方式が提案されている．本チュートリアルでは，WaveNet の登場から最先端のリアルタイムニューラルボコーダまでを紹介し，合成精度，合成速度，モデルサイズ，学習難易度，学習時間等の観点からの比較を行う．

[Tutorial] A comparison of neural vocoders directly synthesizing raw speech waveforms

1. はじめに：直接波形生成ニューラルネットワーク WaveNet の登場

統計的テキスト音声合成や統計的声質変換においては，テキストや変換前の特徴量を入力とし，音響特徴量を推定する音響モデルと，推定された音響特徴量を音声波形へと変換するボコーダが用いられる．前者は，2013 年からの深層ニューラルネットの導入によりモデル性能の向上が見られたが [1,2]，ボコーダに関しては STRAIGHT [3] や WORLD [4] といったソースフィルタに基づく方式が用いられており，最小位相，フレーム内での周期性の仮定，特徴量分析誤差等により，どうしてもこのボコーダが肉声感を阻む大きな壁となっていた．2016 年 8 月の時点では，著者の観点では声帯振動の位相も考慮した Glottal ボコーダ [5] が一番品質の高い方式であったと考えるが，自然音声とはまだ差があった．また，ニューラルネットを用いた直接波形生成モデルの検討もあったが，ソースフィルタの域は超えず，課題が残っていた [6,7].

[1] 情報通信研究機構
National Institute of Information and Communications Technology (NICT)
[2] 名古屋大学
Nagoya Uniersity
[a)] okamoto@nict.go.jp

そのような中，2016 年 9 月のちょうど Interspeech 2016 と SSW9 の会期中に登場したのが，直接波形生成モデル WaveNet [8] であり，従来方式である波形接続方式や LSTM 型音響モデル+ソースフィルタボコーダによる深層ニューラルネット型音声合成法を凌駕し，自然音声にかなり近いテキスト音声合成を実現し，業界に革命を持たらした．WaveNet は画像生成モデル PixelCNN [9] の音声版であり，これまでの音声業界での常識から大きく外れた枠組みであり，この点も業界に大きな衝撃を与えた．

WaveNet の基本原理は，過去の自身の波形サンプルを入力し，次のサンプルを予測する自己回帰モデルである．WaveNet では，Dilated causal convolution を用いて，効率よく長い時系列の音声波形を入力している．WaveNet では音響モデルは存在せず，言語特徴量から推定した音素継続長と基本周波数とを，言語特徴量ベクトルと結合し補助情報として入力し，条件付けを行う．[*1] また，従来のニューラルネットの誤差計算は最小二乗，すなわちガウス分布を仮定することが多いが，WaveNet では，8 bit μ-law 量子化 [12] を適用し，256 個の離散値の分類問題としている点も斬新である．これにより，任意の分布形状がモデル化で

[*1] ただし，この条件付けのネットワーク構造はパラレル WaveNet [10] や WaveRNN [11] も含めて公開されていない．

きる．さらに，音声の非周期成分を適切に表現するために，生成時は出力確率に基づいたサンプリングにより出力波形値を得る．つまり，周期性が高い部分ではソフトマックス関数の出力の確率分布は尖った形をしており (=サンプリングしてもほぼ決まった値が選ばれる)，非周期性が高い部分では分布はフラットとなる (= どの値が選ばれるかはランダムになる)．これらの原理は，ソースフィルタモデルで問題となっていた課題を全て解決し，高品質な合成を実現している．

WaveNet の登場以降，音声合成や声質変換は直接波形生成モデルへと移行した．著者もこれまでに，以下で紹介するサブバンド WaveNet [13,14]，サブバンド FFTNet [15]，単一正規分布型 FFTNet および WaveRNN [16,17]，WaveGlow ボコーダを用いたリアルタイムニューラルテキスト音声合成 [17,18] 等に関する研究に取り組んできた．以下では，WaveNet 登場からちょうど 3 年が絶つ Interspeech 2019 および SSW10 までの波形生成モデルを紹介し，比較を行う．なお，WaveNet に関する研究動向や以下でも紹介するニューラルボコーダの性能比較は文献 [19,20] でも行われている．

2. ニューラルボコーダの紹介

WaveNet は言語特徴量系列から直接音声波形を生成するモデルであるのに対して，従来のソースフィルタボコーダを WaveNet の枠組みで実現する WaveNet ボコーダが提案された [21]．また，WaveNet が登場してすぐ後に出たのが，SampleRNN [22] であり，WaveNet の畳み込みニューラルネットを再帰的ニューラルネットで実現したモデルである．SampleRNN はテキスト音声合成 Char2Wav [23] に適用した際には，WORLD の音響特徴量で条件付けがされている．これらのモデルが，音響特徴量を音声波形へと変換するニューラルネットワーク，ニューラルボコーダである．これ以降，生成時間の高速化，品質改善，複数話者化等，様々な検討が行われている．

また，ニューラルボコーダの登場に伴い，テキストから音声波形への直接変換を目指す End-to-end 音声合成の研究も盛んに行われている．特に，Tacotron 2 では，双方向再帰的ニューラルネット型 sequence-to-sequence モデルにより，英語テキスト系列からメルスペクトログラムを予測し，自己回帰型 WaveNet ボコーダ [*2] により，メルスペクトログラムから音声波形を得ており，遂に自然音声と区別のつかない品質のテキスト音声合成を実現している [24]．

従来のボコーダはソースフィルタに基づく方式であるため，基本周波数とメルケプストラムを音響特徴量としているが，基本周波数の分析には誤差を伴う．これに対して，メルスペクトログラムは短時間フーリエ変換とメルフィルタバンクから一意に求まる特徴量であり，分析誤差はない．それ故，Tacotron 2 の登場以降，ニューラルボコーダの音響特徴量にはメルスペクトログラムが用いられることが多い．

2.1 位相付与に基づく方式

Tacotron 2 の前身である Tacotron [25] や言語特徴量から直接短時間フーリエ変換の振幅スペクトログラムを推定する方式 [26] では，Griffin-Lim 法 [27] により位相成分を付与し，逆フーリエ変換により音声波形を得ていたが，音質はニューラルボコーダには届かない課題がある．

2.2 フロー型高速生成モデル

パラレル WaveNet: WaveNet や SampleRNN は高品質な合成を可能とするが，一番の問題は生成時間であった．これらは共に自己回帰モデルであるため，1 サンプルずつ巨大なネットワークの順伝搬が必要となり，1 秒の音声を合成するのに GPU を用いても 200 秒もかかってしまう大問題である．しかし，WaveNet の登場からわずか 1 年で登場したのが自己回帰型フローに基づくパラレル WaveNet [10] である．自己回帰型 WaveNet が過去の波形と言語特徴量を入力するのに対して，パラレル WaveNet では白色ノイズと言語特徴量を入力とし，全ての音声サンプルを 1 回の順伝搬で一度に生成する．自己回帰型 WaveNet は Dilated causal convolution であるが，パラレル WaveNet は Noncausal な多段の WaveNet によって構成される．学習は自己回帰型フローに基づき，生徒であるパラレル WaveNet の出力を教師である自己回帰型 WaveNet に入力し，それぞれのモデルの出力のカルバック・ライブラー情報量を最小化するようにパラレル WaveNet を学習する．ただし，損失関数がカルバック・ライブラー情報量のみであると「ささやき声」しか合成できないため，音声波形の単時間フーリエ変換で得られるパワー損失等も加え，自己回帰型 WaveNet と同等の音声品質を保ちつつ，GPU 演算によるリアルタイム生成を実現している．また，従来の WaveNet が 8 bit μ-law 量子化を用いた分類問題であるのに対して，パラレル WaveNet では，自己回帰型教師モデルも含めて混合ロジスティック分布による 16 bit 波形予測を行う回帰問題となっており，波形の品質自体も向上している．[*3] 自己回帰型 WaveNet は 1 サンプルごとにサンプリングを行うのに対して，パラレル WaveNet では最初に入力する白色ノイズにより全ての音声サンプルのサンプリングが同時に行われることになる．

単一正規分布型パラレル WaveNet(ClariNet): 上記のモデルは混合ロジスティック分布に基づくが，自己回帰型フローの際のカルバック・ライブラー情報量が解析的に

[*2] パラレル WaveNet [10] と同様，混合ロジスティック分布による 16 bit 波形予測を用いている．

[*3] 混合ロジスティック分布でないとそもそも自己回帰型フロー (知識蒸留) ができないこともある．

計算できないため，サンプリングにより近似を行っている．つまり，学習時には，生徒モデルへ入力する白色ノイズと，カルバック・ライブラー情報量を計算するための近似と，2回のサンプリングが行われている．この問題を解決するために，単一正規分布に基づくパラレル WaveNet [28] が提案された．このモデルでは，モデルの最終出力は平均と標準偏差の 2 次元ベクトルとなり，連続信号を出力できる．単一正規分布に基づくモデルのカルバック・ライブラー情報量は解析的に解けるため，学習時も合成時もサンプリングは最初の白色雑音入力の 1 回のみであり，混合ロジスティック分布を用いた方式よりも高品質は合成を実現している．単一正規分布型モデルで考えると，周期性の高い部分では出力の標準偏差は小さく (=サンプリングしてもほぼ決まった値が出力される)，非周期性が高い部分では出力の標準偏差は大きくなる (=ランダムな信号が出力される)．このモデルを Deep Voice 3 [29] と連結した ClariNet は，最初はそれぞれをメルスペクトログラムを介して学習するものの，その後連結して学習しているため，テキストから音声波形への完全なる End-to-end テキスト音声合成であると言える．また，敵対的生成ネットーワーク (Generative adversarial network: GAN) 学習 [30] を用いた生徒モデルの精度向上も報告されている [31].

生成フロー型ニューラルボコーダ: パラレル WaveNet は自己回帰型フローに基づく方式であるため，自己回帰型の教師モデルが必要であり，また，パワーの損失関数等も導入する必要があるため，学習が複雑である．これに対して，生成フロー型の高速生成モデルが提案されており，WaveGlow [32] と FloWaveNet [33] がある．これらも，パラレル WaveNet と同様，多段の WaveNet を重ねた変数変換による方式であるが，知識蒸留学習はなく，パラレル生成モデルを直接学習できる．これらは全てが逆演算が可能なニューラルネットで構築されている．学習時は音声波形と音響特徴量を入力し，白色雑音を出力するように学習され，生成時は学習時の逆演算により，白色雑音と音響特徴量を入力し，音声波形を生成できる．最先端の音響モデルと WaveGlow ボコーダにより，高品質なリアルタイムテキスト音声合成が実現できる [17,18,34].

2.3 自己回帰型高速生成モデル

パラレル WaveNet や WaveGlow は全てのサンプルを一度に出力できるが，モデルサイズが大きいため，GPU を用いないとリアルタイムの生成はできない．これに対して，WaveNet や SampleRNN のように自己回帰モデルでありながらも，並列生成する方式やモデルサイズを小さくすることによりリアルタイム生成を実現するモデルの検討も行われている．

サブバンド WaveNet: WaveNet はサンプリング周波数 16 kHz や 24 kHz の信号を直接モデル化しているが，学習の前にマルチレート信号処理を用いて音声信号を帯域ごとの $1/M$ のサンプリング周波数の信号に分解し，それぞれの信号を別々の WaveNet でモデル化し，生成時は合成フィルタにより全帯域信号を復元する方式がサブバンド WaveNet である [13]．並列生成が可能であるため，M 倍の合成速度を実現できる他，サンプリング周波数 48 kHz での合成も可能である [14]．課題としては，生成時のサンプリングが帯域ごとに別々に行われるため，位相が揃わない問題がある．位相を揃えるために一番低い帯域の信号を他の帯域にも入力する方式もサブバンド FFTNet を用いて行われているが，音質には課題が残る [15].

FFTNet: FFTNet [35] は，過去の入力サンプルを前後半分ずつに分け，それぞれに 1×1 の畳み込みニューラルネットがかかり，それを足し合わせ，正規化線形関数 (ReLU) $\rightarrow 1 \times 1$ の畳み込み$\rightarrow$ ReLU 後，またその出力を前後半分ずつに分け，上記を最後の 1 サンプルになるまで繰り返し，最後の全結合層を経て，ソフトマックス関数により次の波形サンプルの出力確率を得る．WaveNet と比較すると，モデルサイズは劇的に小さくなるため，CPU を用いてもリアルタイムに生成が可能であるが，学習時の 0 挿入やノイズを混入させた学習，スペクトル減算等を施さないと音質はあまり良くないため，課題が残る．従来の FFTNet は 8 bit μ-law であるが，混合ロジスティック分布モデルや単一正規分布モデルの適用による 16 bit 推定や連続値推定は可能である [16].

WaveRNN: WaveNet や FFTNet は，自身の過去のサンプルを畳み込みニューラルネットワークにより実現しているのに対して，WaveRNN [11] は，再帰的ニューラルネットの 1 つであるゲート付き回帰型ユニット (Gated Rucurrend Unit: GRU)1 層と後段の全結合層 2 層で自己回帰モデルを実現しているため，ネットワークは非常にコンパクトである．また，16 bit のリニア PCM を圧縮なして推定するために，16 bit を大まかな (course)8 bit(256 階調) とその中をさらに 8 bit で表現し (fine)，それぞれを 2 つのソフトマックスで分類問題として推定している．そのため，1 つのサンプルを推定するのに 2 回の順伝搬とサンプリングを必要とする．それでも，GPU を用いることによりリアルタイム生成が可能である．さらに，スパース WaveRNN では，GRU の行列をスパース表現することにより，演算量を大幅に削減し，モバイル CPU でのリアルタイム化も可能である．その上，音声波形系列を N サンプル間引いた波形に分割し，それぞれを別々の WaveRNN で生成する，サブスケール WaveRNN も実現している．この際，生成時はランダムサンプリングされるため，それぞれの波形の位相を考慮するために，最初に推定した少し未来までの系列を次の系列の推定にも入力する方式を取っている．これにより，少しずつ時間をずらして全系列を並列に生成することにより，ほぼ N 倍の速度での合成を実現してい

る．信号処理的に考えれば，高域通過フィルタなしで間引かれた信号は折り返し歪みが発生しているが，[*4]WaveRNN では，他の系列の信号も入力することにより，再構成した際に折り返し歪みが相殺されるように学習されていることになる．また，WaveRNN においても，2 回の順伝搬とサンプリングを避けつつも 16 bit の信号を推定するために，単一正規分布によるモデル化は可能であり，従来法よりもほぼ 2 倍の合成速度を実現している [17]．さらに，複数話者，複数言語の多人数，多言語コーパスで学習し，任意の話者，言語の音声を合成するユニバーサルニューラルボコーダにも WaveRNN が用いられている [36]．

LPCNet: LPCNet は WaveRNN の応用系であり，WaveNet と同様 8 bit μ-law を採用しているが，量子化誤差を減らすために，音声波形を線形予測分析し，その予測残差を推定している．また，WaveRNN と同様，スパース化することによりモバイル CPU でのリアルタイム合成が可能である [37]．LPCNet は圧縮コーディングのニューラルデコーダとして提案されているが，テキスト音声合成も可能である [38]．

2.4 その他の高速生成モデル

フロー型モデル以外の全てのサンプリングを同時に生成する方式の検討も行われている．

MCNN: 単時間フーリエ変換の振幅スペクトル系列を畳み込みニューラルネットを用いて徐々にアップサンプリングしていき，音声波形を得る方式であり，高音質を実現している [39]．音声合成等で用いられるメルスペクトログラムからの変換の検討は行われていない．

Neural source-filter (NSF): 基本周波数とメルケプストラムを入力とし，音声波形を出力するモデルである．入力された基本周波数に基づく調波信号と白色ノイズを多段の変数変換ネットワークにより音声波形へと変換する．WaveNet ボコーダと同等の音質を実現しつつ，GPU を用いたリアルタイム生成が可能なモデルである [40, 41]．

GELP: メルスペクトログラムを全極フィルタへと変換する信号処理と敵対的生成ネットーワーク学習を用いたネットワークにより線形予測の残差成分を推定し，フィルタリングにより全サンプルの音声波形を出力できる [42]．分析合成は高精度であるが，テキスト音声合成時の精度が課題である．

周期非周期分離に基づく方式: 音声信号を基本周波数からなる周期成分と非周期成分とに分離し，周期成分の 1 チャネルと非周期成分の 24 チャネルを出力するネットワークを敵対的生成ネットーワークにより学習し，生成時はこれらを再構成し，リアルタイムに音声波形を得る [43]．

[*4] 通常のマルチレート信号処理に基づくサブバンド WaveNet では折り返し歪みが生じないように分析フィルタを導入している [13–15]．

2.5 品質改善の試み

ノイズシェーピング: 自己回帰型 WaveNet ボコーダでは，8 bit μ-law エンコードによる量子化誤差および推定誤差を生じる．特に後者は高帯域の歪みを生じさせ，これが音質劣化の原因となる．これらを聴覚的に聞こえなくするために，ノイズシェーピング法が提案されており，時不変フィルタを用いる方式 [44] と時変フィルタ [45, 46] を用いる方式とがある．LPCNet [37] や文献 [5] で用いられる声帯振動を WaveNet で推定する GlotNet [47] も時変フィルタを用いる方式に相当する．時不変フィルタを用いる方式では，学習データの平均メルケプストラムを用いて学習データの音声をフィルタリングし，白色化させた音声を用いて WaveNet を学習する．すると，生成時には推定誤差のレベルが高帯域でも信号より大きくならないため，逆フィルタリングにより高品質な音声を合成できる．時不変フィルタを用いる方式は，8 bit μ-law 型 FFTNet，16 bit 型 WaveRNN や単一正規分布型 WaveNet，FFTNet，および WaveRNN にも非常に有効である [15–17]．

WaveCycleGAN: 従来の統計的音声合成の音響モデルで推定した音響特徴量を用いてソースフィルタボコーダで生成した自然音声とは明らかに区別ができる音声を，敵対的生成ネットーワークを用いて，自然音声と同等の品質の音声に波形レベルで変換できる WaveCycleGAN [48] が提案されている．この方式は，不完全な音響特徴量による品質劣化とソースフィルタボコーダによる品質劣化を同時に回復できるネットワークである．また，最近では低サンプリング周波数の音声を高サンプリング周波数への音声と変換する超解像度化の報告もなされている．

QPNet ボコーダ: ニューラルボコーダの問題の 1 つに，学習データの範囲外の基本周波数の音声は合成できない問題がある．この問題を克服するために，Quasi-periodic WaveNet ボコーダ (QPNet) [49] が提案されている．この方式は，入力される基本周波数の値に応じて Receptive field の長さを動的に変えるように学習，生成を行う．これにより，学習データの範囲外の基本周波数の音声が合成できるようになるだけでなく，効率のよい自己回帰モデルとなるため，少ないモデルサイズで通常の WaveNet と同等の品質を実現できる．また，周期非周期分離に基づく方式でも，学習データの範囲外の基本周波数の音声の合成は可能である [43]．

3. ニューラルボコーダの比較

上記のように様々なニューラルボコーダを紹介してきたが，基本的には合成精度，合成速度，モデルサイズ，学習難易度，学習時間等においてトレードオフの関係があると考える．つまり，それぞれのモデルに一長一短があると考える．以下では，著者が実際にコーディングを行った経験のある WaveNet，FFTNet，WaveRNN，パラレル WaveNet，

WaveGlow 間での比較を簡単に紹介する．[*5]

自己回帰モデルで比較すると，WaveNet や FFTNet は，畳み込みニューラルネットを用いているため，学習時は過去のサンプルを全て参照できるため，全てのサンプル予測の損失を一度に計算できる．一方，WaveRNN は再帰的ニューラルネット構造故，学習時も 1 サンプルずつ入力し，過去の履歴を形成する必要があるため，WaveNet や FFTNet に比べ学習には時間を要する．逆に生成時は，モデルサイズが小さいため，WaveRNN の合成速度の方が速い．モデルサイズで比較すると，WaveNet > FFTNet > WaveRNN となるが，モデルサイズが小さい場合，合成速度は速いが，学習時には工夫が必要である．FFTNet や LPCNet では，推論時の自己回帰入力波形の誤差へのロバスト性を獲得するために，学習時の自己回帰入力波形にノイズを混入しており，これを行わない場合は合成時の音声の品質は大幅に劣化する．著者の検討でも，FFTNet や WaveRNN はノイズシェーピングを用いない場合は劣化した音声しか合成できない [15–17]．一方，WaveNet はモデルサイズが大きいため，このような処理がなくとも高品質な音声を合成できる．音質の観点では，やはりノイズシェーピングを用いた WaveNet が一番高いと言える [16, 17]．

パラレル生成モデルの場合，WaveGlow はパラメータ調整なしでも高品質なモデルを実現できるが，学習には多数のバッチサイズが必要なため複数の GPU が必要であり，かつ，学習には非常に長い時間を要するという課題がある．一方，単一正規分布型パラレル WaveNet は WaveGlow ほどの学習時間はかからないが，自己回帰型教師 WaveNet の精度に依存することや，複数の損失関数間の重み調整が必要なため，学習難易度は高い．

また，WaveGlow と WaveRNN で比較すると，WaveGlow は GPU を用いる必要があるという課題があるが，Python 実装のままでもリアルタイム生成を実現できる．一方，WaveRNN は Pytnoh 実装のままでは GPU を用いてリアルタイム生成はできず，リアルタイム生成を実現するには C++等で GPU カーネルを直接制御するコーディングが必要となる [17]．さらに，スパース WaveRNN や LPCNet のようにモバイル CPU を用いたリアルタイム生成を実現するためにも C++等でのコーディングが必要となる．

4. おわりに

本稿では，Interspeech 2018 および SSW10 までのニューラルボコーダを紹介し，著者が携わってきたモデルについての簡単な比較を行った．現在も既に arXiv 等では新しいモデルもいくつか登場しており，今後もさらに新しいモデルが登場するであろう．最後に，本稿がニューラルボコーダを用いた研究開発に関する一助となれば幸いである．

[*5] 学習時間や合成時間は学習コーパスのデータ量やバッチサイズ，計算機環境等によって異なるため，具体的な値は示していない．

参考文献

[1] Ling, Z.-H., Kang, S.-Y., Zen, H., Senior, A., Schuster, M., Qian, X.-J., Meng, H. and Deng, L.: Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends, *IEEE Signal Process. Mag.*, Vol. 32, No. 3, pp. 35–52 (2015).

[2] Wu, Z., Watts, O. and King, S.: Merlin: An open source neural network speech synthesis system, *Proc. SSW9*, pp. 218–223 (2016).

[3] Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A.: Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds, *Speech Commun.*, Vol. 27, No. 3–4, pp. 187–207 (1999).

[4] Morise, M., Yokomori, F. and Ozawa, K.: 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).

[5] Airaksinen, M., Juvela, L., Bollepalli, B., Yamagishi, J. and Alku, P.: Comparison between STRAIGHT, glottal, and sinusoidal vocoding in statistical parametric speech synthesis, *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, Vol. 26, No. 9, pp. 1658–1670 (2018).

[6] Tokuda, K. and Zen, H.: Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis, *Proc. ICASSP*, pp. 4215–4219 (2015).

[7] Tokuda, K. and Zen, H.: Directly modeling voiced and unvoiced components in speech waveforms by neural networks, *Proc. ICASSP*, pp. 5640–5644 (2016).

[8] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K.: WaveNet: A generative model for raw audio, *Proc. SSW9*, p. 125 (2016).

[9] van den Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A. and Kavukcuoglu, K.: Conditional Image Generation with PixelCNN Decoders, *Proc. NIPS*, pp. 4790–4798 (2016).

[10] van den Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., van den Driessche, G., Lockhart, E., Cobo, L. C., Stimberg, F., Casagrande, N., Grewe, D., Noury, S., Dieleman, S., Elsen, E., Kalchbrenner, N., Zen, H., Graves, A., King, H., Walters, T., Belov, D. and Hassabis, D.: Parallel WaveNet: Fast high-fidelity speech synthesis, *Proc. ICML*, pp. 3915–3923 (2018).

[11] Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., van den Oord, A., Dieleman, S. and Kavukcuoglu, K.: Efficient neural audio synthesis, *Proc. ICML*, pp. 2415–2424 (2018).

[12] 711, I.-T. R. G.: *Pulse Code Modulation (PCM) of voice frequencies* (1988).

[13] Okamoto, T., Tachibana, K., Toda, T., Shiga, Y. and Kawai, H.: Subband WaveNet with overlapped single-sideband filterbanks, *Proc. ASRU*, pp. 698–704 (2017).

[14] Okamoto, T., Tachibana, K., Toda, T., Shiga, Y. and Kawai, H.: An Investigation of subband WaveNet vocoder covering entire audible frequency range with limited acoustic features, *Proc. ICASSP*, pp. 5654–5658 (2018).

[15] Okamoto, T., Toda, T., Shiga, Y. and Kawai, H.: Improving FFTNet vocoder with noise shaping and subband approaches, *Proc. SLT*, pp. 304–311 (2018).

[16] Okamoto, T., Toda, T., Shiga, Y. and Kawai, H.: Investigations of real-time Gaussian FFTNet and parallel WaveNet neural vocoders with simple acoustic features, *Proc. ICASSP*, pp. 7020–7024 (2019).
[17] Okamoto, T., Toda, T., Shiga, Y. and Kawai, H.: Real-time neural text-to-speech with sequence-to-sequence acoustic model and WaveGlow or single Gaussian WaveRNN vocoders, *Proc. Interspeech*, pp. 1308–1312 (2019).
[18] Okamoto, T., Toda, T., Shiga, Y. and Kawai, H.: Tacotron-based acoustic model using phoneme alignment for practical neural text-to-speech systems, *Proc. ASRU* (2019).
[19] Boilard, J., Gournay, P. and Lefebvre, R.: A literature review of WaveNet: Theory, application and optimization, *Proc. 146th Conv. Audio Eng. Soc.* (2019).
[20] Govalkar, P., Fischer, J., Zalkow, F. and Dittmar, C.: A comparison of recent neural vocoders for speech signal reconstruction, *Proc. SSW10*, pp. 7–12 (2019).
[21] Tamamori, A., Hayashi, T., Kobayashi, K., Takeda, K. and Toda, T.: Speaker-dependent WaveNet vocoder, *Proc. Interspeech*, pp. 1118–1122 (2017).
[22] Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., Courville, A. and Bengio, Y.: SampleRNN: An unconditional end-to-end neural audio generation model, *Proc. ICLR* (2017).
[23] Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A. and Bengio, Y.: Char2Wav: End-to-End Speech Synthesis, *Proc. ICLR* (2017).
[24] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y. and Wu, Y.: Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions, *Proc. ICASSP*, pp. 4779–4783 (2018).
[25] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R. and Saurous, R. A.: Tacotron: Towards End-To-End Speech Synthesis, *Proc. Interspeech*, pp. 4006–4010 (2017).
[26] Takaki, S., Kameoka, H. and Yamagishi, J.: Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis, *Proc. Interspeech*, pp. 1128–1132 (2017).
[27] Griffin, D. and Lim, J. S.: Signal estimation from modified short-time Fourier transform, *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 32, No. 2, pp. 236–243 (1984).
[28] Ping, W., Peng, K. and Chen, J.: ClariNet: Parallel wave generation in end-to-end text-to-speech, *Proc. ICLR* (2019).
[29] Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J. and Miller, J.: Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning, *Proc. ICLR* (2018).
[30] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative adversarial nets, *Proc. NIPS*, pp. 2672–2680 (2014).
[31] Yamamoto, R., Song, E. and Kim, J. M.: Probability density distillation with generative adversarial networks for high-quality parallel waveform generation, *Proc. Interspeech*, pp. 699–703 (2019).
[32] Prenger, R., Valle, R. and Catanzaro, B.: WaveGlow: A flow-based generative network for speech synthesis, *Proc. ICASSP*, pp. 3617–3621 (2019).
[33] Kim, S., Lee, S.-G., Song, J., Kim, J. and Yoon, S.: FloWaveNet : A generative flow for raw audio, *Proc. ICML*, pp. 3370–3378 (2019).
[34] Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z. and Liu, T.-Y.: FastSpeech: fast, robust and controllable text to speech, *Proc. NeurIPS* (2019).
[35] Jin, Z., Finkelstein, A., Mysore, G. J. and Lu, J.: FFTNet: A real-time speaker-dependent neural vocoder, *Proc. ICASSP*, pp. 2251–2255 (2018).
[36] Lorenzo-Trueba, J., Drugman, T., Latorre, J., Merritt, T., Putrycz, B., Barra-Chicote, R., Moinet, A. and Aggarwal, V.: Towards achieving robust universal neural vocoding, *Proc. Interspeech*, pp. 181–185 (2019).
[37] Valin, J.-M. and Skoglund, J.: LPCNet: Improving neural speech synthesis through linear prediction, *Proc. ICASSP*, pp. 5826–7830 (2019).
[38] Kons, Z., Shechtman, S., Sorin, A., Rabinovitz, C. and Hoory, R.: High quality, lightweight and adaptable TTS using LPCNet, *Proc. Interspeech*, pp. 176–180 (2019).
[39] Arik, S. O., Jun, H. and Diamos, G.: Fast Spectrogram Inversion Using Multi-Head Convolutional Neural Networks, *IEEE Signal Process. Lett.*, Vol. 26, No. 1, pp. 94–98 (2019).
[40] Wang, X., Takaki, S. and Yamagishi, J.: Neural source-filter-based waveform model for statistical parametric speech synthesis, *Proc. ICASSP*, pp. 5916–5920 (2019).
[41] Wang, X. and Yamagishi, J.: Neural harmonic-plus-noise waveform model with trainable maximum voice frequency for text-to-speech synthesis, *Proc. SSW10*, pp. 1–6 (2019).
[42] Juvela, L., Bollepalli, B., Yamagishi, J. and Alku, P.: GELP: GAN-excited linear prediction for speech synthesis from mel-spectrogram, *Proc. Interspeech*, pp. 694–698 (2019).
[43] Oura, K., Nakamura, K., Hashimoto, K., Nankaku, Y. and Tokuda, K.: Deep neural network based real-time speech vocoder with periodic and aperiodic inputs, *Proc. SSW10*, pp. 13–18 (2019).
[44] Tachibana, K., Toda, T., Shiga, Y. and Kawai, H.: An investigation of noise shaping with perceptual weighting for WaveNet-based speech generation, *Proc. ICASSP*, pp. 5664–5668 (2018).
[45] Yoshimura, T., Hashimoto, K., Oura, K., Nankaku, Y. and Tokuda, K.: Mel-cepstrum-based quantization noise shaping applied to neural-network-based speech waveform synthesis, *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, Vol. 26, No. 7, pp. 1173–1180 (2018).
[46] Song, E., Byun, K. and Kang, H.-G.: ExcitNet vocoder: A neural excitation model for parametric speech synthesis systems, *Proc. EUSIPCO* (2019).
[47] Juvela, L., Bollepalli, B., Tsiaras, V. and Alku, P.: GlotNet—A raw waveform model for the glottal excitation in statistical parametric speech synthesis, *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, Vol. 27, No. 6, pp. 1019–1030 (2019).
[48] Tanaka, K., Kaneko, T., Hojo, N. and Kameoka, H.: Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks, *Proc. SLT*, pp. 632–639 (2018).
[49] Wu, Y.-C., Hayashi, T., Tobing, P. L., Kobayashi, K. and Toda, T.: Quasi-periodic WaveNet vocoder: A pitch dependent dilated convolution model for parametric speech generation, *Proc. Interspeech*, pp. 196–200 (2019).