一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2022-15(2022-06)

[招待講演] Crazy vocoder は砕けない
～でもちょっとくだけた未来の話を～

森勢　将雅†

† 明治大学総合数理学部　〒164–8525 東京都中野区中野 4–21–1
E-mail: †mmorise@meiji.ac.jp

あらまし　現在の音声合成研究者が論文に Vocoder と記載するとき，その多くは Deep neural network (DNN) を用いて何らかのパラメータから高品質な音声波形を生成する Neural vocoder を指すのではないだろうか．もしそうであれば，音声符号化という役割ではなく，高品質な音声を合成したいという高品質 Vocoder が持つ『黄金の精神』が，Neural vocoder 世代に受け継がれたことを意味する．本稿では，恐らく今後失われていくであろう伝統的な Vocoder の波形生成部のアルゴリズム，および一連の知識がまだ音声研究において役立つかという将来展望について紹介する．
キーワード　音声合成，Vocoder，テキスト音声合成，声質変換，音声デザイン

[Invited talk] Crazy vocoder is unbreakable
—But let's talk about an informal vision of the future—

Masanori MORISE†

† School of Interdisciplinary Mathematical Sciences, Meiji University
4–21–1 Nakno, Nakno-ku, Tokyo, 164–8525 Japan
E-mail: †mmorise@meiji.ac.jp

Abstract　When current speech synthesis researchers refer to Vocoder in their papers, they are most likely referring to Neural vocoder, which generates high-quality speech from parameters by using deep neural networks (DNN). If so, the "golden spirit" of a high-quality vocoder, which is to synthesize high-quality speech rather than play the role of speech encoding, has been passed on to the Neural vocoder generation. This paper presents the core algorithms in the waveform generation of traditional vocoder, which will probably be lost in the future, and prospects for how this body of knowledge can still be useful in speech research.
Key words　Speech synthesis, vocoder, text-to-speech synthesis, voice conversion, speech design

1. ま え が き

「音声合成」とは，辞書的には計算機により人工的に音声を生成する技術のことと定義されており，その定義からも幅広い研究領域をカバーしていることは想像に難しくない．本稿は音声合成を構成する要素技術の 1 つである Vocoder [1] を話の中心に据えるため，まずは音声合成に関する区分と現状を説明するところから出発する．

2020 年代に入り，計算機により合成された音声は日常生活でも耳にするようになった．最先端の技術で合成された合成音の品質は，既に肉声と区別できない水準に達している [2]．スマートスピーカや駅のアナウンス等で使われる合成音声を生成する技術は，音声合成の中でもテキスト音声合成（Text-To-Speech; TTS）と呼ばれ，主要なカテゴリの 1 つといえるだろう．TTS の場合，入力はテキスト情報，出力は音声波形となるが，例えば日本語では漢字を含むテキストの解析が必要であり，テキスト解析は現在のところ別モジュールで処理している．したがって，入力から出力までに複数のモジュールを通す必要があり，日本語特有の工夫も必要である [3]．近年では，アカデミア以外のニュース等で合成音声について取り上げられるようになり，TTS のことを指して音声合成と表現することもあるが，音声合成の研究は TTS 以外にも複数のカテゴリが存在する．TTS については，実現する技術にもいくつかのパラダイムシフトがあったため，本稿では統計的パラメトリック音声合成 [4] 以降の技術を指して TTS という用語を用いる．

声質変換（Voice conversion）は，TTS と並んで音声合成にお

ける主要な分野である．声質変換をボイスチェンジャーと考えると，ある人の音声を別人の音声に変換する技術を想像するかもしれないが，声質変換そのものは音声に含まれる言語以外の特徴を変換することと定義される [5]．そのため，他人の音声に変換する技術は，話者変換や個人性変換等と呼ばれ，声質変換に包含される技術としてカテゴライズされる．上記のような高度な加工だけではなく，物理量として定義可能な高さや音色の加工も，言語情報を変化させない場合は声質変換といえるだろう．なお，変調スペクトル法 [6] のように高さや音色などのパラメータを用いない方法も存在するため，本稿で Vocoder により得られる音声パラメータを用いるものに対象を絞る．

TTS，声質変換のどちらも，2020 年代に入り音声パラメータを用いない方法が増えつつあるが，未だに音声パラメータを用いた研究も進められている．本稿のタイトルにもある Vocoder は，広義では音声から音声を構成するパラメータを推定し，推定されたパラメータ群から音声波形を生成する形式と定義できる．初期の Vocoder は，品質を優先していないため合成音らしさが顕著であったが，STRAIGHT [7] に代表される高品質な Vocoder は，TTS や声質変換の基盤を支える技術として用いられてきた．Vocoder は，パラメータ群から音声波形を生成するというモジュールが存在するため，音声合成の枠組みに含まれる．他の研究と区別する場合は，音声分析合成 [8] という用語が用いられる．

これらの分野において，高品質な Vocoder は重要な役割を担ってきたが，その役割は Deep neural network (DNN) を用いた Neural vocoder に継承されつつある．Neural vocoder は DNN により音声パラメータから波形を生成する機能を有し，名称が最初に提案されたのは Char2Wav [9] であると考えられる．本稿では，Vocoder における波形生成に注目し，高品質 Vocoder の内容と問題点，今後の展望について紹介する．

2. Vocoder に期待する特長の変化と現状

本稿において Vocoder と記載した場合，1939 年に Dudley により発表された，Channel vocoder [1] の原理に基づくものを指すこととする．Vocoder の名称を含むものでは Phase vocoder [10] も知られているが，分析により得られるパラメータが異なるため，本稿では対象としない．以下では，Vocoder が発表後から現代に至るまでどのように進化を遂げたのかについて説明する．

2.1 音声符号化のための Vocoder

Vocoder は Voice と Coder を合わせた言葉であるように，元々は符号化のための技術である．音声通信において効率よく情報を伝達することが重視されており，音質よりも優先して，発話内容を限られたビットレートで通信することが目的とされてきた．基本的な考え方では，図 1 に示すソース・フィルタ理論に基づいて，声帯振動に相当するパルス列（ソース）と口の形状等による音色付けを実現するフィルタの畳み込みで音声を表現する．連続する声帯振動をパルス列 $x(t)$，音色付けを実現するフィルタを $h(t)$ とすれば，音声波形 $y(t)$ は

$$y(t) = x(t) * h(t), \tag{1}$$

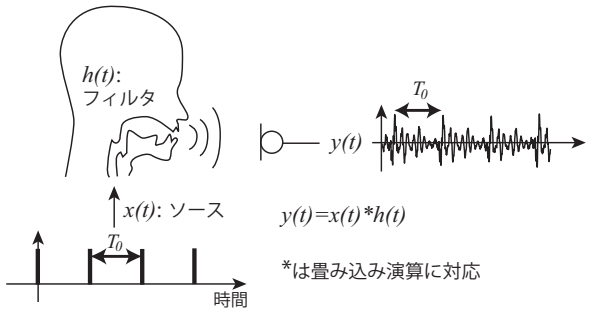

図 1　ソース・フィルタモデルの考え方．基本周期 T_0 の区間毎にパルスを生成し，スペクトル包絡のインパルス応答 $h(t)$ と畳み込むことで波形を得る．スペクトル包絡には，声帯振動，口の形状，口元からの放射特性が包含されることに注意する．

と，ソース情報とフィルタ情報との畳み込みにより表される．

パルス列 $x(t)$ における各パルスの間隔は，基本周波数に基づいて決定する．Vocoder におけるパラメータでは，$h(t)$ のスペクトルである $H(\omega)$ が用いられており，このスペクトルから言語情報を優先して残すように圧縮することで，効率の良い音声伝送を実現する．基本周波数は 1 フレームについて 1 つの数値で表現される一方，スペクトル包絡は 1 フレームについて多次元の情報を有するため，主にスペクトル包絡の圧縮が重要となる．伝統的には線形予測符号（Linear predivtive coding; LPC）を用いた LPC vocoder [11] や，ケプストラム [12] を用いた準同型 Vocoder [13] が提案されてきた．

Vocoder の音質が悪い理由はいくつかあるが，ソース・フィルタ情報の推定精度以外にも，非周期的な成分が考慮されないことが挙げられる．Vocoder では有声音は周期的な成分であるパルス列で駆動させているが，人間の音声には有声音中にも非周期的な成分である雑音が混入している．非周期的な成分の有無が知覚に与える影響については 1960 年代から指摘されており [14]，従来の Vocoder を拡張して非周期的な成分を扱うための方法も検討されてきた．具体的には，Mixed source model [15] や，Multiband excitation vocoder [16] が該当する．

Vocoder は，1990 年代半ばまでは符号化効率を優先するという条件下で品質を高めるというアプローチが主流であった．符号化効率を優先した当時の Vocoder の音質は明らかに悪いとされてきたが，これには，符号化を重視したアルゴリズムを基盤として改善してきたことも，原因の 1 つと考えられる．1990 年代に，符号化効率という縛りを考慮せずに提案された STRAIGHT [7] が提案されたことにより，Vocoder は高品質 Vocoder として認識が改められることになる．

2.2 高品質 Vocoder

STRAIGHT は，非周期的成分を扱うためのパラメータを含めた 3 種の音声パラメータを入力音声から推定するアルゴリズム，3 種の音声パラメータから音声波形を生成するアルゴリズムから構成される．本稿では，図 2 のように入力音声から音声パラメータを推定する部分を Encoder，音声パラメータから波形を生成する部分を Decoder として Vocoder の構成要素を定義することにする．STRAIGHT が提案されたことで，Vocoder に

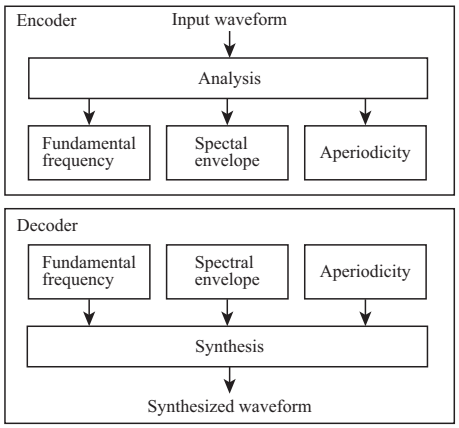

図 2　3 種類の音声パラメータに基づく高品質 Vocoder．波形から 3 種の音声パラメータ（基本周波数：Fundamental frequency，スペクトル包絡：Spectral envelope，非周期性指標：Aperiodicity）を取り出す Encoder と，3 種の音声パラメータから波形を生成する Decoder から構成される．

は肉声に匹敵する音声を合成できるポテンシャルがあることが示された．一方，STRAIGHT により得られる音声パラメータのビットレートは，波形そのもののビットレートよりも高い．このことは，Vocoder という名称であるが，目的が符号化ではなく，高品質な音声合成や柔軟な音声加工に目的がシフトしたことを意味する．同時に，音声パラメータを加工することで品質劣化を抑えつつ声質を加工できる基盤技術としての役割が生じることとなった．声質変換の 1 つである音声のモーフィング [17] も，STRAIGHT を基盤として提案された新たな変換法の 1 つである．

高品質 Vocoder が様々な音声処理の基盤として利用されるようになり，前述の TTS の研究の基盤にも STRAIGHT が使われてきた．現在の主流は STRAIGHT から筆者が提案した WORLD [18] (D4C edition [19]) に変わりつつあるが，これは WORLD が特許を取得せずオープンソースで公開していることが理由と考えられる．最新版の WORLD は，基本周波数推定には Harvest [20]，スペクトル包絡推定には CheapTrick [21], [22]，非周期性指標推定には D4C [19] が採用されている[1]．

高品質 Vocoder は他にも検討されており，例えば Log-domain pulse model [24] が挙げられる．音声をパラメータで表現する方法は Phase vocoder をはじめ複数あり，目的に応じて多様な方法が利用されている．

2.3　Neural vocoder の実現

特に TTS において Vocoder が重要であった主な理由は，テキスト情報から波形を直接出力することが困難であったことに由来する．2010 年代前半までは HMM による統計的パラメトリック音声合成が主流であり，2013 年に DNN 音声合成 [25] が提案されているが，どちらも Vocoder の利用を前提に，音声パラメータを出力としていた．その後，GlottDNN [26] のように有声音の励起信号に相当する部分を DNN で生成するアプローチが提案されてきた．2016 年に提案された WaveNet [27] は，Vocoder の特徴量を経由せずに波形を生成できる方法であり，それまでの DNN 音声合成と比較して一線を画した品質改善を成した．

音声波形を直接出力できることが確認された以降，DNN による波形生成技術が提案されるようになった．その先駆けとして Char2Wav [9] と WaveNet vocoder [28] が 2017 年に提案されている．この時点での Neural vocoder は，Vocoder のパラメータから波形を生成する，Vocoder における decoder の代替手段であった．Tacotron 2 [2] では Wavenet vocoder を変形し，メルスペクトログラムから波形を生成するようにしており，肉声と有意な差が検出できない品質を達成した．現状では，Neural vocoder は Vocoder の音声パラメータに限らず音響特徴量から波形を生成する Neural network のこととして用いられ，TTS における主流は Neural vocoder に置き換えられつつある．DNN による TTS 全般については，文献 [29] でまとめられている．

このように Neural vocoder は，音声符号化の意識ではなく，人間と等価な品質で音声を合成したいという高品質 Vocoder のコンセプトが受け継がれ，合成結果の品質に比重が置かれていると言える．一方 Tacotron 2 ではサンプリング周波数が 24 kHz で実現していることや，DNN を用いるためリアルタイム処理が困難であることなどが，従来の Vocoder に対するデメリットであった．Neural vocoder の研究は現在も盛んであり，2021 年には LPCNet [30] に基づくフルバンド LPCNet [31] が提案されることで，これらの問題も解決されつつある．

2.4　Vocoder における Decoder の問題点

Vocoder により合成された音声の品質が Neural vocoder に劣ることには，いくつかの原因が考えられる．なお，ソース・フィルタモデルにおける基本周波数とスペクトル包絡に非周期的な成分を扱うパラメータを用いた合成では，Vocine [32] や PLATINUM [33] などが提案されている．本稿では，WaveNet vocoder などと対比した議論を展開するため，非周期的な成分を非周期性指標として扱い合成する方法を対象とする．WaveNet vocoder [28] の論文では，STRAIGHT と比較しているが波形生成が MLSA のため，STRAIGHT 本来の波形生成アルゴリズムとの比較はできない．ただし，WaveNet との比較はされており有意な差が検出できないということと，WaveNet が既存の TTS よりも高品質であったことから，Vocoder のパラメータから波形生成する DNN は既存の Vocoder より優れていると仮定できる．なお，Tacotron 2 のようにメルスペクトログラムを入力する場合は，基本周波数の推定誤差や有声無声判定のミスによる影響は生じないなど，問題設定に差があるため本稿での比較の対象とはしない．

ここからは，音声分析合成 [8] を参考に，WORLD で採用している波形生成処理を説明しつつ，現状の Vocoder の問題点について述べる．第一の問題は，非周期性指標の存在にある．非周期的な成分が存在している場合，ソース・フィルタモデルには以下のように非周期性成分 $n(t)$ が加算されることになる．

$$y(t) = x(t) * h(t) + n(t), \tag{2}$$

Vocoder で推定されるスペクトル包絡は $h(t)$ の振幅スペクトル

(注1)：初期バージョンではスペクトル包絡推定が STAR [23] であるなど，現在のバージョンに至るまで数度のアルゴリズム変更がなされている

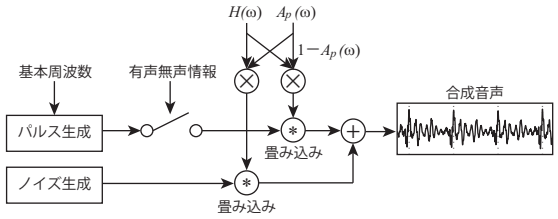

図 3　音声パラメータ群から波形を生成する図の流れ（書籍 [8] の 2 章を一部改変して引用）.

$|H(\omega)|$ であるが，スペクトル包絡には非周期性成分の影響が含まれることとなる．非周期性指標は，推定された振幅スペクトル $|\hat{H}(\omega)|$ のうち非周期性成分が占める割合を各周波数に対して算出する音声パラメータである．$y(t)$ の振幅スペクトルにおいては，周期性成分と非周期性成分のスペクトルが複素数で加算されているため，同位相であれば振幅は加算され，逆位相であれば振幅は減算される．よって，推定された振幅スペクトルに対して成分の分離を行うアプローチは，厳密な成分分離が出来ているとは言い難く，この影響が合成結果にも繋がる可能性が存在する．

Vocoder における波形生成では，基本周波数軌跡 $f_0(t)$ から声帯振動に相当するパルスを生成する時刻を算出する処理が含まれる．WORLD や STRAIGHT では，可能な限り厳密な時刻を算出するため，軌跡に対し以下の処理を実施する．まず，基本周波数軌跡 $f_0(t)$ から以下の式により回転角度に相当する情報を算出する．

$$\theta(t) = 2\pi \int_0^t f_0(\tau) d\tau, \tag{3}$$

ここで，例えば基本周波数が 100 Hz に固定される場合は，10 ms で $\theta(t)$ が 2π 変化することになる．つまり，$\theta(t)$ を 2π 変化する区間が基本周期と一致する．$\theta(t)$ を 2π で割った余りを算出し，前のフレームとの変化量が負値の場合にパルスが生じたと判断することが可能となる．一連のパラメータの例を示した結果が図 4 である．

このような処理でパルスが生じる時刻を算出しているが，これでは時刻が標本化周波数に依存したサンプル点でのみ生じることとなる．そのため，以下のようにスペクトルにおいて任意の遅延時刻 τ に基づく成分 $e^{-i\tau\omega}$ を乗じて逆フーリエ変換することで，サンプル点に縛られないパルスの励起を実現する．

$$x(t-\tau) = \int_{-\infty}^{\infty} e^{-i\tau\omega} X(\omega) e^{i\omega t} d\omega, \tag{4}$$

この処理は，ナイキスト周波数周辺において強いパワーを持つ場合では時間波形に影響が生じるが，音声の声帯振動のパワーが低域で支配的であることや，AD 変換においてその周辺のパワーがアンチエイリアシングフィルタで抑制していることから，品質に与える影響は少ないと考えている．本処理には全ての時刻において 0 より大きい基本周波数が必要であるため，WORLD では無声音の部分の基本周波数を 500 Hz に設定している．これは，無声音のスペクトルを 2 ms 毎に変化させることを可能にするため，主に破裂音の合成において有利に働くことを期待した数値である．

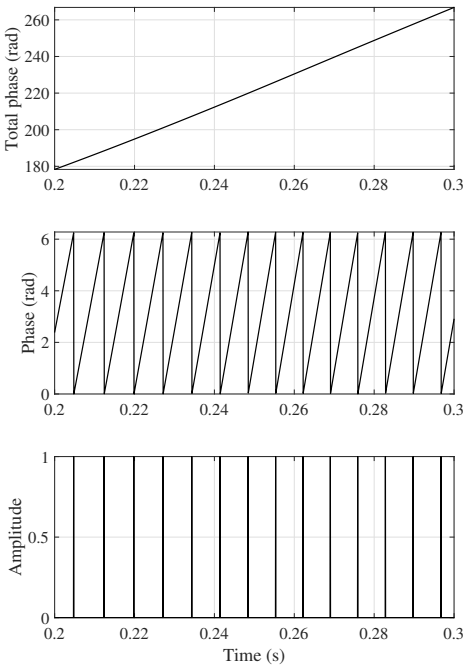

図 4　Vocoder におけるパルス励起時刻の算出方法．上段は基本周波数の積分により求めた位相．中段は，位相を 2π で割った余り．下段は，中段で位相ジャンプに基づき生成したパルス列.

このように，パルスの生成時刻算出についてはかなり厳密な処理をしているが，第二の問題はパルスに畳み込むスペクトルの情報である．スペクトル包絡は振幅値であり位相情報を持たないため，STRAIGHT や WORLD では最小位相を算出して利用している．最小位相はゼロ位相よりも品質が高いとされているが，最大位相成分が無視されているため，その分劣化が生じることが予測される．STRAIGHT では高域の群遅延を操作することでこの問題に対処しており，WORLD では特段の処理を入れていない．両者の品質の差は軽微であるが，最大位相を無視しており，Vocoder 固有の「Buzzy」とされるブザー音に近い音色の劣化が生じている．WaveNet 関連の技術を用いた波形生成においてこの Buzzy さが生じないことは，波形の位相成分を良好に推定できている可能性を示唆する．

3.　Vocoder のポテンシャルと将来展望

Vocoder は Neural vocoder と比較して合成結果の品質は劣るが，本稿を執筆している 2022 年時点ではまだ Neural vocoder では実現できていない機能があり，利用価値も残されている．本節では，現時点で Vocoder が役に立つと思われる研究事例と，将来展望について紹介する．

3.1　Vocoder による音源解析

TTS などで利用される Neural vocoder は，Vocoder の名前はついているものの TTS や声質変換のモジュールが出力した音声パラメータ（メルスペクトログラム等を含む）を入力とするため，Decode 機能に特化していると言える．Vocoder における Encoder 部分は，例えば楽器音解析 [34] における演奏パフォー

マンスを比較する特徴量の推定に利用されている．つまり，音声や調波構造を有する楽音の解析を目的とした研究において，元音源に近い波形を出力できる精度のパラメータが出力できる機能には需要があると言える．

3.2 音声知覚における音声生成

Neural vocoder の現時点での問題は，学習データに存在しないパラメータに対し生成される音声の品質の予測が困難なことである．Vocoder であれば，例えば全体のピッチを一定量シフトさせる処理やスペクトル包絡を伸縮する処理などを，厳密に実装することが可能になる．これは，主に実験心理学分野において音声知覚の実験デザインをする際に強力な性質となる．Vocoder は利用していないが，ピッチを加工した音声による心理学実験 [35] のように，音声のパラメータを系統的に変化させて人間の知覚がどのように変化するかを観測する実験において，Vocoder は有効に機能する．他にも，内田が 2016 年に発見した音高錯覚 [36] のように，ピッチとスペクトル包絡の伸縮を精密に制御することで知覚するピッチが基本周波数と逆転する現象なども，Vocoder の存在により発見されたといえるだろう．

3.3 TTS における精密なパラメータ制御

End-to-End 方式における TTS は，入力情報がテキスト情報に限定されるため，基本周波数を精密に操作する機能と相性が良いとは言い難い．筆者らが提案した Human-in-the-loop 音声合成 [37] では，Vocoder のパラメータをユーザが操作し，その結果を入力として自然な音声を出力する技術について検討している．しかしながら，前節で述べたように学習データに存在しないパラメータでは明確な品質劣化が生じるため，高品質と柔軟な加工の両立は現状では困難と言っても良いだろう．

人間と等価な品質が実現できる End-to-End TTS であるが，現在は特徴を制御する Controllable speech synthesis の研究も進められている．FastSpeech [38] では話速の調整ができる TTS であり，Emotion-controllable speech synthesis [39] のように感情をターゲットとした方法も提案されている．これらの方法は，現時点では発話全体に対するパラメータとして入力できることであり，例えば基本周波数軌跡をフリーハンドで制御した結果から自然な音声を合成するような技術は困難である．

3.4 将 来 展 望

Vocoder は TTS という領域において役割を終えつつあるが，まだ有効活用できる領域は残されている．とはいえ，将来的に Vocoder は Neural vocoder に駆逐されるのか，あるいは何らかの形で共存するのかについては，今後の研究次第と言えるだろう．下手な展望を書くと将来黒歴史になるリスクはあるが，ボンクラとして語り継がれるのもまた一興として，今後の方向性について無責任に述べることとする．

現在の End-to-End 音声合成は，スマートスピーカのようにユーザが細かく加工することを想定しない用途において，すでに中心的な存在になりつつある．DNN を利用した TTS はスマートフォン等での利用は困難であると思われるが，これはスマートフォンの性能向上に加え，省メモリ・低コスト化の研究によりクリアされるだろう．つまり，スマートスピーカなどに用いる加工を前提としない TTS では，Vocoder は近い将来 Neural vocoder に駆逐される可能性が高いと考えている．

コンテンツ制作において合成音声は現在も利用されており，昨今では製品からフリーソフトまで幅広い選択肢がクリエイタに提供されている．この領域では音声を精密に制御してコンテンツを「作り込む」作業が行われており，そこでは人間の音声に近いことは必ずしも作り込みが不要であることを意味しない．前述のように，End-to-End TTS においてもパラメータが加工な音声合成が進められる可能性はあるが，多数のパラメータを精密に制御するリアルタイム技術の実現はまだ先であると予想している．このような用途では，Vocoder のように人間の音声からやや品質が劣化したとしても，パラメータに忠実な音が生成される性質が有効であると言える．

音声パラメータからの波形生成が可能な WaveNet vocoder であれば上述の問題をクリアできるが，これは Vocoder と共存可能であろう．例えば，Vocoder によるリアルタイム制御で音声パラメータを加工してイメージを作り，最終的に品質の良い音声を WaveNet vocoder でレンダリングするような使い方が想定される．また，Neural vocoder は学習データに品質が依存するため，任意の話者に対する合成という用途では Vocoder のほうが相性が良い．不特定多数の音声に対して動作するリアルタイム処理が可能な Neural vocoder が得られた際には，Vocoder の市場は縮小していく可能性は高い．

心理学分野のように，音声に対して知覚する感性情報と物理量を紐づける研究においては，Vocoder は今後も利用されることが想定される．これは，パラメータに忠実な音声が生成されることは実験デザインにおける条件の 1 つだからである．このような研究では Vocoder は今後も有効であることから，Vocoder における波形生成の問題点を信号処理で解決する価値はあると考えられる．原理的には Vocoder のパラメータから波形を生成する Neural vocoder と同等の品質は達成できるはずであり，そのための課題を扱うことは引き続き検討していきたい．

4. お わ り に

Vocoder は，人間と同じ品質の音声を合成するという TTS のタスクにおいては役割を終えつつある．一方で，心理学分野における音声知覚の実験において人間を計測するための加工ツールを考えた場合，音声を柔軟に加工できるという性質を持つ Vocoder にはまだ役割が残されている．将来的には，Vocoder が持つメリットをも包含する Neural vocoder が提案される可能性も否定はできない．それらが気軽に動作できる計算コストで実現できるようなったときこそが，真に Vocoder が役割を終えたときと考えている．

現時点では，いわゆるソース・フィルタモデルで近似できる音声が主な対象であり，声帯振動が規則的に生じない特殊な音声の合成は未だ実現が困難なタスクである．すなわち，現状の音声合成は，人間が発声できる極一部というローカルなゴールを達成しただけといえる．人間が発声可能なあらゆる音声を計算機で生成することや，人間が音声をどのように知覚しているかのメカニズム解明等，多くの課題が残されている．それらの研究に着手するためのツールとして，手軽に利用できる Vocoder

にはまだ利用価値があると考えている．今後も DNN を用いずに動作する Vocoder と，その品質を Neural vocoder に近づけるための研究を進めていきたいと考えている．

謝　　辞

本研究の一部は，JSPS 科研費 JP21H04900，JP21K19794 の支援を受けました．また，論文中で紹介した WORLD から始まる一連のプロジェクトは，多くの研究者とユーザに支えられてきました．関係する皆様に感謝申し上げます．

文　　献

[1] H. Dudley, “Remaking speech,” J. Acoust. Soc. Am., vol. 11, no. 2, pp. 169–177 (1939).

[2] J. Shen, R. Pang, R.J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R.A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, “Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions,” in Proc. ICASSP2018, pp. 4779–4783 (2018).

[3] 山本龍一，高道慎之介，“Python で学ぶ音声合成，” 株式会社インプレス，2021.

[4] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” Speech Communication, vol. 51, no. 11, pp. 1039–1064 (2009).

[5] T. Toda, A. W. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2222–2235 (2007).

[6] K. Kobayashi, T. Toda, and S. Nakamura, “Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential,” Speech Communication, vol. 99, pp. 211–220 (2018).

[7] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction,” Speech Communication, vol. 27, no. 3–4, pp. 187–207 (1999).

[8] 森勢将雅，“音声分析合成，” コロナ社 (2018).

[9] J. Sotelo, S. Mehri, K. Kumar, J. Felipe Sntos, K. Kastner, A. Courville, and Y. Bengio, “Char2Wav: End-To-End speech synthesis,” in Proc. ICLR 2017, pp. 1–6 (2017).

[10] J. L. Flanagan and R. M. Golden, “Phase vocoder,” The Bell System Technical Journal, vol. 45, no. 9, pp. 1493–1509 (1966).

[11] B. S. Atal and S. L. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” J. Acoust. Soc. Am., vol. 50, no. 2, pp. 637–655 (1971).

[12] A. Noll, “Short-time spectrum and “cepstrum” techniques for vocal pitch detection,” J. Acoust. Soc. Am., vol. 36, no. 2, pp. 269–302 (1964).

[13] A. V. Oppenheim, “Speech analysis-synthesis system based on homomorphic filtering,” J. Acoust. Soc. Am., vol. 45, no. 2, pp. 458–465 (1969).

[14] O. Fujimura, “An approximation to voice aperiodicity,” IEEE Trans. on Audio and Electroacoust., vol. 16, no. 1, pp. 68–72 (1968).

[15] J. Makhoul, R. Viswanathan, R. Schwartz, and A. Huggins, “A mixed source model for speech compression and synthesis,” J. Acoust. Soc. Am., vol. 64, no. 6, pp. 1577–1581 (1978).

[16] D. W. Griffin and J. S. Lim, “Multiband excitation vocoder,” IEEE Trans. on Acoust., Speech, and Signal Process., vol. 36, no. 8, pp. 1223–1235 (1988).

[17] H. Kawahara, M. Morise, H. Banno, and V. G. Skuk, “Temporally variable multi-aspect N-way morphing based on interference-free speech representations,” in Proc. APSIPA ASC 2013, pp. 1-10 (2013).

[18] 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. E99-D, no. 7, pp. 1877–1884 (2016).

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

[20] M. Morise, “Harvest: A high-performance fundamental frequency estimator from speech signals,” in Proc. INTERSPEECH 2017, pp. 2321–2325 (2017).

[21] M. Morise, “CheapTrick, a spectral envelope estimator for high-quality speech synthesis,” Speech Communication, vol. 67, pp. 1–7 (2015).

[22] M. Morise, “Error evaluation of an F0-adaptive spectral envelope estimator in robustness against the additive noise and F0 error,” IEICE transactions on information and systems, vol. E98-D, no. 7, pp. 1405–1408 (2015).

[23] 森勢将雅，松原貴司，中野皓太，西浦敬信，“高品質音声合成を目的とした母音の高速スペクトル包絡推定法，” 電子情報通信学会 論文誌 D，vol. J94-D, no. 7, pp. 1079–1087 (2011).

[24] G. Degottex, P. Lanchantin, and M. Gales, “A log domain pulse model for parametric speech synthesis,” IEEE/ACM Trans. on Audio, Speech, and Langage Process., vol. 26, no. 1, pp. 57–70 (2018).

[25] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in Proc. ICASSP 2013, pp. 7962–7966 (2013).

[26] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku, “GlottDNN — A full-band glottal vocoder for statistical parametric speech synthesis,” in Proc. INTERSPEECH 2016, pp. 2473–2477 (2016).

[27] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” arXiv preprint arXiv:1609.03499 (2016).

[28] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, “Speaker-dependent WaveNet vocoder,” in Proc. INTERSPEECH 2017, pp. 1118–1122 (2017).

[29] 高木信二，“話声の合成における応用技術，” 日本音響学会論文誌，vol. 75, no. 7, pp. 393–399 (2019).

[30] J. Valin and J. Skoglund, “LPCNet: Improving neural speech synthesis through linear prediction,” in Proc. ICASSP 2019, pp. 5891–5895 (2019).

[31] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, Y. Shiga, and H. Kawai, “Full-Band LPCNet: A Real-Time Neural Vocoder for 48 kHz Audio With a CPU,” IEEE Access, vol. 9, pp. 94923–94933 (2021).

[32] Y. Agiomyrgiannakis, “Vocaine the vocoder and applications in speech synthesis,” in Proc. ICASSP 2015, pp. 4230–4234 (2015).

[33] M. Morise, “PLATINUM: A method to extract excitation signals for voice synthesis system,” Acoust. Sci. & Tech., vol. 33, no. 2, pp. 123–125 (2012).

[34] A. Lee, S. Furuya, M. Morise, P. Iltius, and E. Altenmüller, “Quantification of instability of tone production in embouchure dystonia,” Parkinsonism & Related Disorders, vol. 20, no. 11, pp. 1161–1164 (2014).

[35] B. C. Jones, D. R. Feinberg, L. M. DeBruine, A. C. Little, and J. Vukovic, “Integrating cues of social interest and voice pitch in men’s preferences for women’s voices,” Biology letters, vol. 4, pp. 192–194 (2007).

[36] T. Uchida, “Reversal of the relation between impressions of voice pitch and height of fundamental frequency: Cognitive biases caused by conversion of tone quality,” Acoust. Sci. & Tech., vol. 39, no. 2, pp. 143–146 (2018).

[37] D. Kondo and M. Morise, “Human-in-the-loop speech-design system and its evaluation,” in Proc. APSIPA ASC 2019, pp. 608–612 (2019).

[38] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, “FastSpeech: fast, robust and controllable text to speech,” in Proc. NIPS’19, pp. 3171–3180 (2019).

[39] X. Luo, S. Takamichi, T. Koriyama, Y. Saito and H. Saruwatari, “Emotion-controllable speech synthesis using emotion soft labels and fine-grained prosody factors,” in Proc. APSIPA ASC 2021, pp. 794–799 (2021).