2-2-6　ニューラル音声合成技術
2-2-6　Neural Speech Synthesis Technology

岡本 拓磨
OKAMOTO Takuma

テキストから自然な音声波形を合成するテキスト音声合成（Text-to-speech synthesis: TTS）技術は、ニューラルネットワークを用いた方式の進展により、現在では CPU のみを用いて高品質かつリアルタイムな合成が可能となっている。NICT においても、音声翻訳アプリ VoiceTra® において、最先端の技術を用いた多言語ニューラル音声合成を導入している。本稿では、入力テキストから音響特徴量を推定するニューラル TTS 音響モデル、音響特徴量から音声波形を生成するニューラル音声波形生成モデル、また、音響特徴量を介さず 1 つのニューラルネットワークを用いてテキストから音声波形を直接生成可能である End-to-end モデル及び NICT における取組について紹介する。

Text-to-speech synthesis (TTS) technology, which synthesizes natural speech waveforms from input texts, can now realize high-quality synthesis in real-time only using CPUs, thanks to recent advances in neural network-based methods. NICT has also introduced multilingual neural speech synthesis using state-of-the-art technologies in VoiceTra®, a speech translation application for smartphones. This paper briefly introduces neural TTS acoustic models which predict acoustic features from input texts, neural speech waveform generative models which synthesize speech waveforms from acoustic features, end-to-end TTS models which directly synthesize speech waveforms from input texts without intermediate acoustic features, and related neural TTS models that NICT has been working on.

1　はじめに

入力テキストを機械が自然な音声で読み上げるテキスト音声合成（Text-to-speech synthesis: TTS）は、音声コミュニケーションにおいて重要な技術の 1 つである。近年では、自動音声ガイダンス、駅等での自動アナウンス、スマートスピーカ、カーナビ、対話ロボット、等の日常の様々な場面で使われるようになり、TTS は身近な技術となっている。NICT においても、言語の壁を超えた音声コミュニケーションの実現に向けて、本特集でも紹介されている音声認識技術と機械翻訳技術を音声合成技術と組み合わせることにより、多言語音声翻訳を実現し、VoiceTra® 等において幅広く利用されている。そして現在は、入力音声を即座に別言語の音声へと変換する同時通訳システムの研究開発に取り組んでいる。同時通訳システムにおいても、高品質かつ高速な音声合成技術は重要な研究課題の 1 つである。

2012 年頃までは隠れマルコフモデル（Hidden Markov model: HMM）に基づく統計的音声合成 [1] が主流であり、NICT においても HMM を用いた多言語音声合成技術の開発に取り組んできた [2]。この方式で合成された音声は、話している内容は問題なく聞き取ることができるが、人間の自然音声と比べると明らかに自然性に乏しいもの（いわゆるロボットのような声）であり、大きな課題となっていた。

その中で、2012 年頃から様々な分野において飛躍的な技術革新を持たらしたのが深層ニューラルネットワーク（Deep neural network: DNN）であり、現在のいわゆる AI ブームが始まった。音声認識や機械翻訳と同様、ニューラルネットを用いた方式は 2013 年に Google によって音声合成にも導入され、HMM に基づく方式を上回る精度を実現した [3][4]。NICT においても、ニューラルネットを用いた方式が検討され [5]、2022 年 3 月まで VoiceTra® にも搭載されていた。しかし当時は、ニューラルネットが導入されたのは図 1（a）における音響モデルのみであり、最終的な音声波形生成部は既存の信号処理に基づく方式を採用し

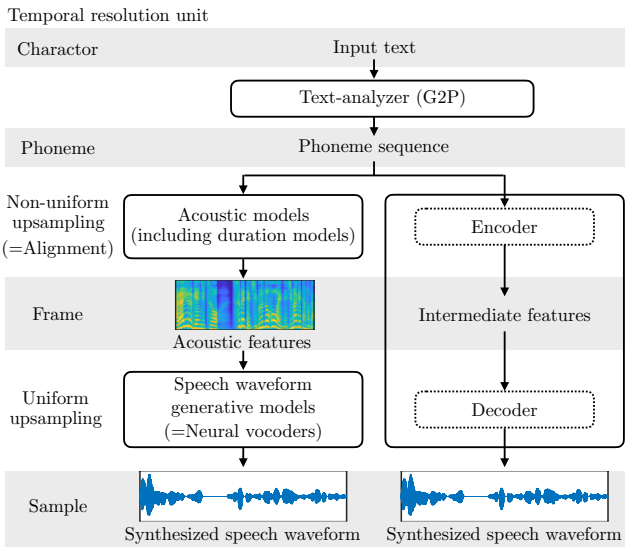
ていたため、自然音声の品質にはまだ届かず、課題解決には至らなかった。

しかし、その課題を解決に導いたのが、2016 年 9 月に Google DeepMind から突如提案された音声波形生成ニューラルネットモデル WaveNet [6] であり、翌年に提案された Tacotron 2 [7] により、TTS においてついに自然音声と同等の品質を達成した。そこから非常に数多くのモデルが提案され、現在では CPU のみを用いて高品質かつリアルタイムな TTS が可能となっている。NICT においても、2022 年 3 月に、最先端のニューラルネットワーク技術を用いた日英中韓越 5 言語のニューラル TTS を VoiceTra® に搭載した。VoiceTra® では、現在合計 19 言語の多言語 TTS が利用可能であり、日英中韓越以外の言語についても順次ニューラル TTS へと更新予定である。

本稿では、10 年前の報告 [2] と同様、TTS のみに焦点を当て、ある話者の声を別の話者の声に変換する声質変換（Voice conversion: VC）[8] 技術については割愛する。まず、**2** において TTS における音声波形生成問題の難しさについて述べる。次に、**3** で WaveNet を紹介した上で、**4** において入力テキストから音響特徴量を推定するニューラル TTS 音響モデルを解説し、**5** にて音響特徴量から音声波形を生成するニューラル音声波形生成モデル及び音響特徴量を介さず 1 つのニューラルネットワークを用いてテキストから音声波形を直接生成可能である End-to-end モデルを紹介する。また、**4** 及び **5** において、NICT における取組（筆者らの査読論文 [9]–[21]）について適宜紹介する。最後に、**6** にてまとめと今後の課題について述べる。なお、紙面の都合上、参考文献は主要なもののみを引用している。また、**5** のニューラル波形生成モデルの詳細については著者による解説記事 [22] を参照されたい。

2　TTS における音声波形生成問題の難しさ

従来の TTS では、入力テキストと出力音声間の中間表現として、音声波形を短時間フレームごとに周波数分析した音響特徴量が用いられる。図 1（a）に示すとおり、入力文がテキスト解析により音素（+ アクセント）系列へと変換され、音響モデルにより音響特徴量へと変換される。ここで、各音素が音響特徴量の何フレーム分に相当するかは、音素アライメントによる音素継続長モデルや注意機構モデルにより推定され、変換時は各音素のフレーム数に応じた不均一なアップサンプリングにより、時間解像度が音素単位からフレーム単位となる。次に、音声波形生成モデルにより、音響特徴量から音声波形を生成する。ここで、音響特徴量は音声波形から固定のフレームシフト量で分析するため、均一なアップサンプリングにより、時間解像度がフレーム単位からサンプル単位へと変換される。一方、End-to-end モデルでは、陽な音響特徴量は介さないものの、エンコーダからフレーム単位の中間特徴量を生成し、デコーダにより音声波 形を生成する（図 1（b））。つまり、現状の TTS では、音素から中間特徴量への不均一なアップサンプリング及び中間特徴量から音声波形への均一なアップサンプリングの 2 段階の異なるアップサンプリングにより、テキストから音声波形への変換を実現している。



(a) Pipeline TTS models　(b) Entire end-to-end TTS models

図 1　ニューラルネットワークを用いたテキスト音声合成モデル。(a): パイプラインモデル、(b):End-to-end モデル

例えば、英単語「hello」の場合、アルファベットではたった 5 文字であるが、音声波形になると、仮に長さ 1.0 秒としても、サンプリング周波数 24 kHz、フレームシフト量 12.5 ms の場合は、80 フレーム、24,000 サンプルにも及ぶ。通常、人間は全く同じ発話は二度と発声できないため、聴感上は同じであっても、フレーム単位、サンプル単位では発話ごとに毎回異なる。つまり、TTS における音声波形生成問題とは、入力された系列長の数千倍以上（フレーム単位からは数百倍）の出力系列を「確率的」に求める極めて難しい問題である。また、ニューラルネットを含む機械学習における回帰問題では、学習データと出力結果間の平均二乗誤差（Mean square error : MSE）損失を最小化するようにモデルを学習するが、音声信号は波形信号ではあるが非周期成分（= ランダム性）も多く含まれているため、MSE 損失では標準正規分布の平均値しか推定できず（= 非周期成分は平均化されてしまい精度よく生成できない）、音質に大きな影響を与える。

そのため、従来は多数の音声波形を短い素片成分へと分割し、それらをつなぎ変えて波形を生成する素片

接続方式 [23] や、HMM 型 TTS や DNN 型 TTS では、ソースフィルタ理論に基づき、音響特徴量を基本周波数（= 声帯振動に対応）、スペクトル包絡（= 声道形状に対応）及び非周期成分とし、信号処理を用いて波形生成を行うソースフィルタボコーダ（STRAIGHT [24]、WORLD [25] 等）が用いられてきた。

しかし、素片接続方式については要求される音声データの分量と接続部の音声劣化、ソースフィルタボコーダについては最小位相及びフレーム内の周期性の仮定や特徴量分析等が、それぞれ肉声感を阻む大きな要因となり、DNN 音響モデル [3][4] によって精度の高い音響特徴量が推定できたとしても、高品質な合成には至らなかった。

そこへ 2016 年 9 月に突如登場し、TTS や VC に革命をもたらし、肉声感のある音声合成を実現したのが WaveNet [6] である。

3　音声波形生成モデル：WaveNet

WaveNet [6] は、過去の音声波形サンプル $x_0,\cdots,x_{t-1}$ を入力とし、フレーム単位にアップサンプリングしたテキスト解析結果である言語特徴量 h で条件付けした場合の、時刻 t の音声波形の条件付き出力確率 $p(x_t\,|x_0,\cdots,x_{t-1},h)$ を出力するニューラルネットである（図 2 (a)）。ここで、過去の音声波形の周期パターン、非周期性や微細構造等の特徴を効果的に捉えるために、多段の因果的な Dilated convolutional neural network（CNN）を用いている。さらに、音声波形に 8 bit μ-law 量子化を適用し、MSE 損失最小化のような回帰問題ではなく、256 階調の分類問題として交差エントロピー損失を最小化するようにモデルを学習している。これにより、正規分布ではなく、任意の確率分布形状をモデル化できる。そして、音声の周期性及び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 x_t を得る。つまり、入力された過去の波形及び言語特徴量から WaveNet が次のサンプルは周期性が高いと推定した部分では出力される確率分布 $p(x_t\,|x_0,\cdots,x_{t-1},h)$ 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）、逆に非周期性が高いと推定した部分では確率分布 $p(x_t\,|x_0,\cdots,x_{t-1},h)$ はフラットとなる（= どの値が選ばれるかはランダム）。これらの原理により、**2** で述べた従来の素片接続方式及びソースフィルタボコーダの問題点を解決し、肉声感のある高品質な合成を実現した。

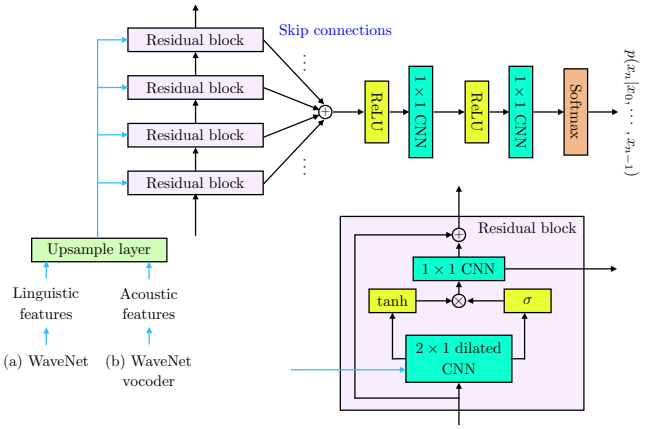


図 2　(a): WaveNet、(b): WaveNet ボコーダ

WaveNet の成功を受けて、言語特徴量ではなく、ソースフィルタボコーダの音響特徴量で条件付けされた WaveNet ボコーダ [26]（図 2 (b)）が提案され、同じくソースフィルタボコーダを上回る品質を実現した。これ以降、数多くの「ニューラルボコーダ」が登場し、TTS、VC 及び歌声合成等において、ニューラルボコーダが使われるようになった。

WaveNet は自己回帰モデルであるため、過去の波形情報を入力として使える分推定問題としては容易となり高品質な合成を実現できるが、生成時間がリアルタイムとは程遠いという課題があった（1 秒の音声を合成するのに GPU を用いても 200 秒）。しかし、WaveNet の登場からわずか 1 年で、白色雑音と言語特徴量を入力すると全てのサンプルを同時に生成可能な Parallel WaveNet [27] が提案され、リアルタイムな高速生成が可能となった（**5. 2** 参照）。

NICT においても WaveNet の高品質生成能力に着目し、WaveNet の登場初期から検討を行い、ノイズシェーピング [28] 及びサブバンド WaveNet [9]–[11] を提案した。前者ついては、WaveNet の誤差分布は通常周波数上にフラットに広がるが、高域は音声のパワーが小さいため誤差が目立ちやすくなる問題に対して、ノイズシェーピングによりあらかじめ高域のスペクトルを持ち上げた音声で学習することにより、知覚的な音質劣化を低減できる。ノイズシェーピングは WaveNet 以外の自己回帰型音声波形生成モデルにおいても有効である [11][12]。後者のサブバンド化については、図 3 に示すとおり、音声波形をマルチレート信号処理によって複数帯域の信号に分割し、帯域ごとに WaveNet を学習、生成することにより生成速度を向上させることができる方式である。また、複数話者で学習した WaveNet ボコーダに時間伸縮した音響特徴量を入力することにより、学習に用いていない任意の話者に対するニューラル話速変換を提案し、従来の信号処理に基づく方式よりも高品質な変換を実現できることを示した [18]。

そして、Sequence-to-sequence 型 TTS モデル Taco-

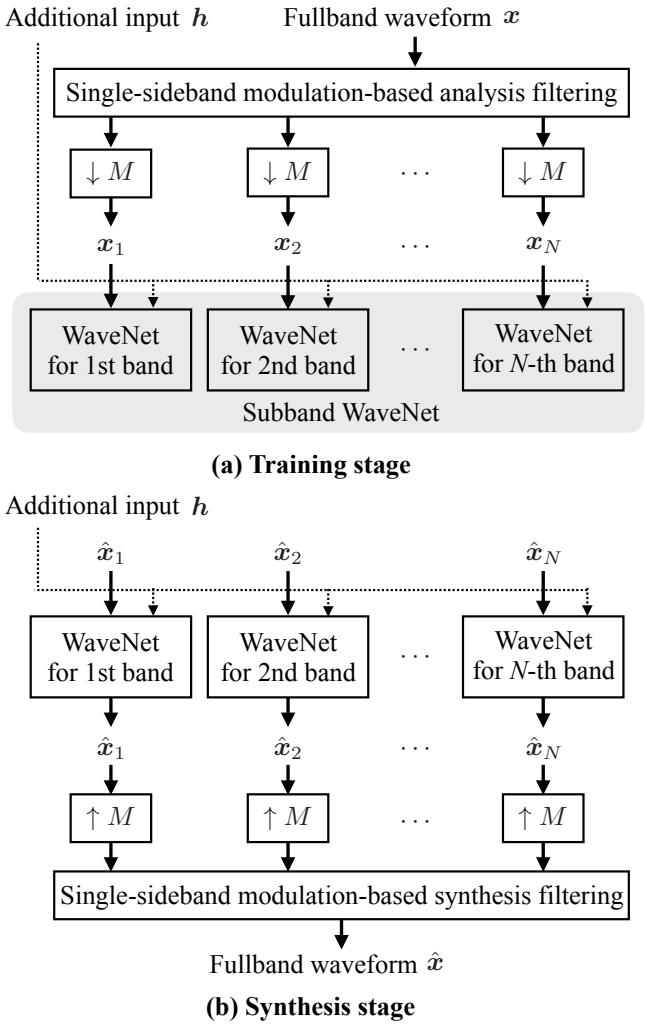


図 3　サブバンド WaveNet

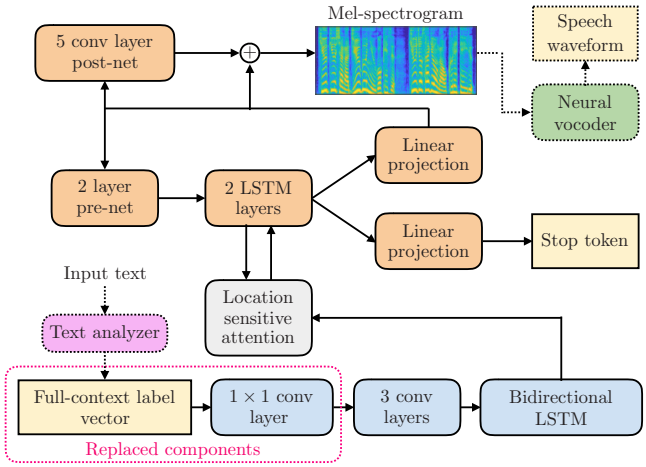


図 4　フルコンテキストラベル入力型 Tacotron 2

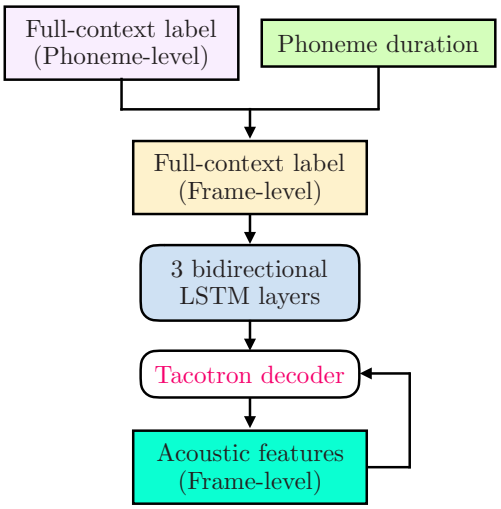


図 5　音素継続長入力型 sequence-to-sequence 音響モデル

tron 2 [7] が提案され、図 1 (a) の音響特徴量をメルスペクトログラムとして WaveNet ボコーダへ入力(= 条件付け)することにより、英語テキスト入力の TTS において、ついに人間と同等の合成品質を実現した。

以下では、WaveNet 及び Tacotron 2 登場以降のニューラルネットに基づく TTS 音響モデル及び音声波形生成モデルの急速な進展について紹介する。

4　ニューラル TTS 音響モデル

4.1　自己回帰型モデル

これまでの HMM 音響モデル [1] や DNN 音響モデル [3][4] では、テキストと音響特徴量間の時刻対応付けである音素アライメントが必要であり、HMM 等を用いて別途外部アライメントモデルを学習する必要があった。また、フレーム単位の学習であるため、言語特徴量には前後の音素やアクセント等の情報を含める必要があった。これに対して、Tacotron 2 では、Sequence-to-sequence モデルとして発話単位の学習となるため、前後の音素の情報は不要となり、また、ニューラル機械翻訳で提案された注意機構を導入することにより、外部アライメントモデルなしで直接アライメントの推定が可能となった [7]。Tacotron 2 は再帰的ニューラルネットを用いているため学習が遅いという問題に対して、同じくニューラル機械翻訳で提案された Transformer を用いた音響モデルも提案され、Tacotron 2 と同等の高品質合成を実現した [29]。

NICT における取組として、日本語のようなピッチアクセント言語に対応した Tacotron 2 として、図 4 に示すテキスト解析結果であるフルコンテキストラベル入力型モデルを提案し、後述する WaveGlow ボコーダ [30] と組み合わせることにより、高品質かつ GPU を用いたリアルタイム生成が可能な日本語ニューラル TTS を実現した [13]。さらに、注意機構は外部アライメントは不要であるが、まれに推定時にアライメント予測に失敗し、発話が途中で止まる、スキップされる音素や繰り返し発話される音素を生じる致命的な問題

に対して、Tacotron 2 や Transformer に既存の外部アライメントを組み込んだ高品質かつ安定したニューラル TTS を提案した（図 5）[14][15]。

4.2　非自己回帰型モデル

Tacotron 2 や Transformer はニューラル音声波形生成モデルと組み合わせることにより非常に高品質な TTS を実現できるが、上述のとおり、注意機構の予測失敗による発話の破綻、失敗という実サービスにおいては致命的な問題を有する。また、注意機構に基づくモデルは次のフレームの出力を得るために過去の出力を入力とする自己回帰モデルであるため、生成速度が遅いという課題があった。これらの問題を解決するために、非自己回帰 Sequence-to-sequence 型ニューラル TTS 音響モデルである FastSpeech [31] が提案され、安定かつ高速な TTS を実現した。FastSpeech では、教師モデルとして学習した Transformer の音素アライメントを用いて音素継続長モデルを学習し、自己注意型ネットワークを用いたエンコーダ出力を音素継続長に応じてアップサンプリングを行い、同じく自己注意型ネットワークを用いたデコーダで音響特徴量を出力する。さらに、教師モデルや外部アライメントを必要とせず、Soft-DTW を用いて音素アライメントを自動で獲得しつつ、自己回帰型モデルと同等の品質を実現する Parallel Tacotron 2 [32] などが提案されている。

NICT の取組としては、外部音素アライメントモデルを導入した安定して学習可能な Parallel Tacotron 2 を実装し、後述する Multi-stream HiFi-GAN と組み合わせて、CPU のみでリアルタイム生成可能な高品質 TTS を実現した [17]。

5　ニューラル音声波形生成モデル

WaveNet の成功と課題を受けて、肉声感のある高品質を保ちつつ、リアルタイム高速生成可能なニューラル音声波形生成モデルが数多く提案された。いずれも、画像生成分野等で開発された深層生成モデルを音声波形生成へと移植したモデルとなっており、以下の 4 種類に大別される。以下では、それぞれのモデルの特徴及び NICT における取組について紹介する。ここで、自己回帰モデル（図 6 (a)）以外は全ての音声波形サンプルを一度に生成するパラレル生成モデル（図 6 (b)-(e)）であり、高速生成を実現できるが、過去の波形情報を使えないため、推定問題としては難しくなる。しかし、**5.3** で紹介する HiFi-GAN [33] 等の高尚な深層波形生成モデルでは、高速生成かつ自己回帰モデルを超える高音質を実現している。

5.1　高速型自己回帰型モデル

WaveNet は非常に巨大なネットワーク構造であるためリアルタイム生成できない問題に対して、自己回帰モデルではあるがネットワークが軽量であるため、CPU のみで高品質かつリアルタイム生成を実現可能な WaveRNN [34] や LPCNet [35] が提案されている。しかし、これらの自己回帰モデルは、特に TTS や VC で推定された「鈍った」音響特徴量を用いた場合、ごくまれに突然波形がクリップし、爆音を発する「Collapsed speech [36]」を生じるため、実サービスでの実装ではこの問題を解決しておくことが必須である。

NICT での取組としては、LPCNet は 1 時間程度の音声データで学習可能であることを示し [19]、また、人間の可聴域をカバーするサンプリング周波数 48 kHz の音声合成を可能とする Full-band LPCNet を提案している [20]。

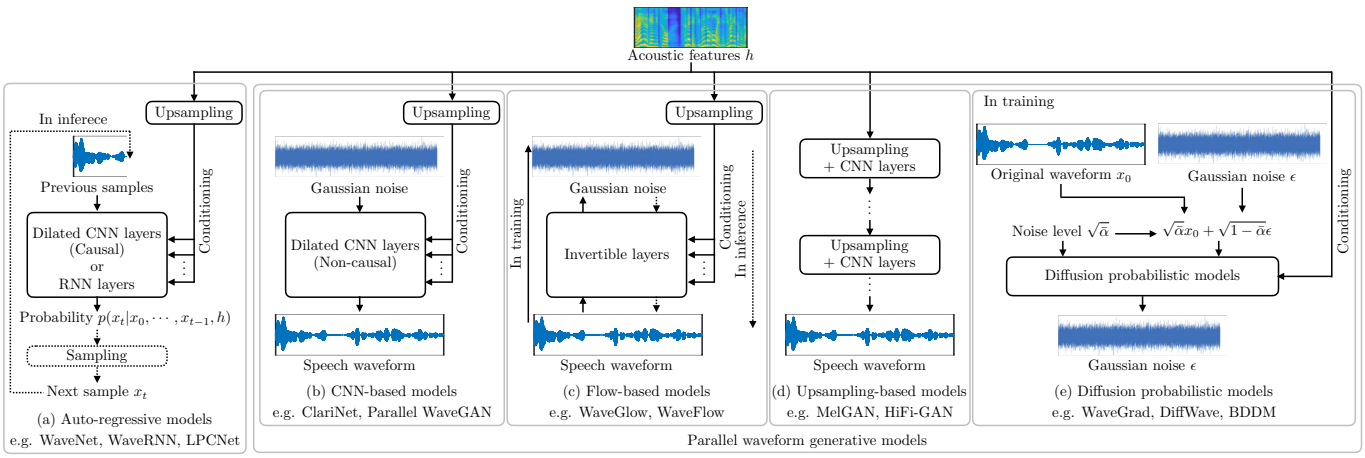


図 6　ニューラルネットワークを用いた音声波形生成モデル。(a): 自己回帰モデル、(b): 畳み込みニューラルネット型パラレル生成モデル、(c):Flow 型パラレル生成モデル、(d): アップサンプリング型パラレル生成モデル、(e): 拡散確率型パラレル生成モデル

5.2　Flow 型生成モデル

WaveNet の登場からわずか 1 年で提案された高品質高速生成モデル Parallel WaveNet [27] は、Inverse autoregressive flow（IAF）[37] に基づき、教師モデルとしての自己回帰モデルから生徒であるパラレル生成モデルを知識蒸留により学習する。 また、知識蒸留における教師モデルと生徒モデル間のカルバック・ライブラー情報量を解析的に算出可能な WaveNet の出力確率を単一正規分布（分類問題ではなく標準正規分布の平均と分散を推定する回帰問題）とした ClariNet [38] も提案されている。具体的には、生徒であるパラレル生成モデル（図 6（b））に白色雑音とメルスペクトログラム（音響特徴量）を入力し、音声波形を出力する。出力した音声波形を教師である自己回帰モデル（図 6（a））へ入力し、教師モデルの出力と生徒モデルの出力間のカルバック・ライブラー情報量を最小化するように生徒モデルを学習する。これにより、教師モデルと出力確率が一致するように学習されるため、自己回帰型 WaveNet と同等の音声品質を保ちつつ、GPU を用いたリアルタイム生成を実現している。

NICT における取組として、メルスペクトログラムではなく、ソースフィルタボコーダ用の音響特徴量を用いた ClariNet の検討 [12] や、単一正規分布型 WaveRNN [13] を提案している。

Parallel WaveNet や ClariNet は自己回帰型の教師モデルが必要であるのに対して、Flow 型生成モデル [39] に基づく WaveGlow が提案された [30]。WaveGlow は全てが逆演算可能なニューラルネットであるため、教師モデルを必要とせず、パラレル波形生成モデルを直接学習できる。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（図 6（c））。Flow 型生成モデルでは、学習時の損失関数は最終出力である白色雑音の負の対数尤度（ゆうど）及び、各変数変換におけるヤコビアンの総和によりシンプルに与えられ、損失を十分小さくできれば、白色雑音から任意の変換が可能であるため、微細構造や非周期成分も精度よくモデル化・生成できる。しかしこれらモデルは、**5.3** の敵対的生成モデルと比較した場合、高精度な変換を実現するためには巨大なネットワーク、学習時間及び十分なデータ量が必要となる課題がある。

NICT における取組として、フルコンテキストラベル入力型 Tacotron 2 や Transformer と組み合わせた GPU を用いた日本語リアルタイム TTS を検討した [13]–[15]。

5.3　敵対的生成モデル

敵対的生成モデル（Generative adversarial network: GAN）[40] は、実際に信号を生成する生成器と、生成器を訓練するための識別器の 2 つのニューラルネットを同時に学習する。ここで、GAN においては、生成モデルの確率分布は陽に定めず、生成器は識別器を騙（だま）すように学習される。逆に、識別器は生成器に騙されない（= 原信号と生成信号とを見分ける）ように学習される。これら 2 つのモデルを同時に「敵対的」に学習させることにより、お互いのモデル精度を向上させる。生成器と 識別器に十分な表現能力があり、かつ学習データも十分である場合は、生成器はデータの真の生成確率を獲得できることが理 論的に示されている [40]。つまり、敵対的生成モデルでは、いかに緻密な生成器、識別器を設計・学習するかが鍵となる。敵対的生成モデルに基づく音声波形生成モデルは数多く提案されているが、以下では、現在最も広く使われている HiFi-GAN [33] について述べる。

HiFi-GAN [33] は、入力されたフレーム単位の音響特徴量に対して、数段のアップサンプリング層と畳み込みにより、白色雑音の入力なしに直接サンプル単位の音声波形を得る（図 6（d））。HiFi-GAN では、生成器に Multi-receptive field fusion という異なるカーネルサイズ、dilation サイズの複数の CNN による出力を統合した畳み込み層を導入することにより、異なる長さの波形パターンを表現できるようになり、Parallel WaveNet 等と比べると段数の少ない CNN でありながら、高精度かつ高速な変換を実現している。また、Multi-period discriminator と Multi-scale discriminator という 2 つの識別器を導入することにより、音声波形の周期パターン及び連続性や長期依存性をそれぞれモデル化している。HiFi-GAN では、これらの洗練されたネットワークにより、高速かつ高品質な音声合成を実現している [33]。これらのモデルは、白色雑音を入力していない（= サンプリングしない）ため、同じ特徴量では毎回同じ波形を出力する。サンプリングなしで高品質な合成を実現できるのは、生成器が音響特徴量に対する適切な微細構造及び非周期成分を「コピー」し、識別器に見破られないようにそれらを適切に「ペースト」して出力しているためであると考えられる。ClariNet や他のパラレル生成モデルはリアルタイム生成のためには GPU が必要であるのに対して、HiFi-GAN は CPU のみで高品質かつリアルタイム生成が可能なパラレルモデルであり、かつ公式実装が公開されていることもあり、現在最も広く使われているニューラル音声波形生成モデルである。

NICT の取組としては、HiFi-GAN 生成器における最後の 4 倍のアップサンプリングをゼロ挿入型アップ

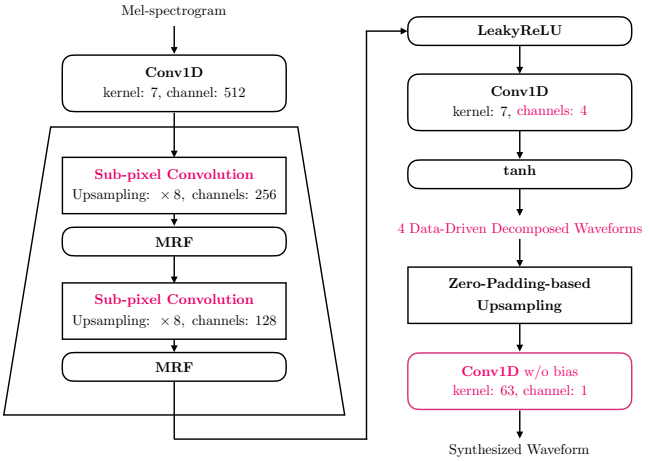


図 7　Multi-stream HiFi-GAN 生成器

サンプリングと学習可能な CNN に置き換えた Multi-stream HiFi-GAN を提案し（図 7）、合成品質を保ちつつ、合成速度を向上させた。さらに、外部アライメント型 Parallel Tacotron 2 と組み合わせることにより、CPU のみでリアルタイム生成可能な高品質ニューラル TTS を実装した [17]。また、メルスペクトログラムではなく、ソースフィルタボコーダ用の低次元な音響特徴量を用いた場合でも高品質な合成が可能であることを示した [21]。

5.4　拡散確率モデル

拡散確率モデルとは、入力信号に徐々に白色雑音を加えていくと（ステップ 0 → N）最終的には白色雑音となる拡散過程に対して、その逆変換であるステップ n から $n-1$ 間の雑音除去過程を学習するモデルである（図 8）。画像生成において提案された深層生成モデルであるが [41]、すぐさま音声波形生成モデル WaveGrad [42] と DiffWave [43] が提案された。拡散確率型波形生成モデルでは、入力した白色雑音に対して、雑音除去と少しずつレベルを下げた白色雑音の加算とを交互に繰り返し、徐々に音声波形へと変換する。そのために、音声波形 x_0 と白色雑音 ε とを重み付きで重畳した信号を入力とし、重畳した雑音 ε のみを推定するモデル ε_θ を学習する（図 6 (e)）。合成時は、入力した白色雑音 $x_N \sim N(0, I)$ が雑音除去過程により徐々に音声波形へと変換される（$n = N \rightarrow 1$）。WaveGrad 及び DiffWave は上記の学習及び合成アルゴリズムにより実現される。WaveGrad は数段のアップサンプリング・ダウンサンプリング層により実現され、DiffWave は Parallel WaveNet [27] 等で広く用いられている図 6 (b) のような非因果的な多段 Dilated CNN を採用している。

拡散確率モデルは、他の深層生成モデルと異なり、時間信号領域での単純な MSE [43]（または L1 [42]）損失のみで学習できる。N を 1,000 等の非常に大きな値とすれば原音に匹敵する高音質が得られるが、生成時間がリアルタイムとは程遠いため、N を 10 以下等に小さくしても高音質を実現できる雑音スケジュール β_n をいかに設定するかが課題となる。

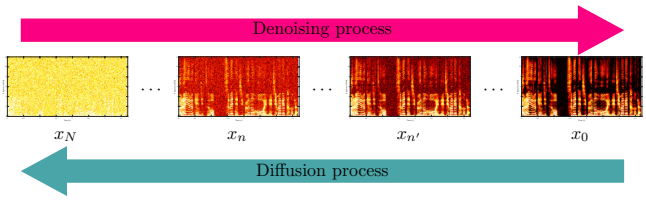


図 8　拡散確率モデルにおける拡散過程（左向き方向）及び雑音除去過程（右向き方向）

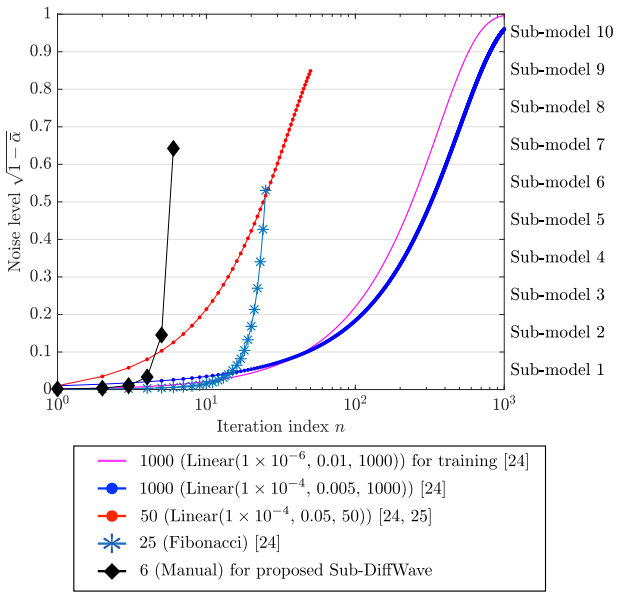


図 9　拡散確率モデルにおける雑音レベル分割型サブモデリング

NICT での取組として、雑音レベルごとに異なるモデルを学習するサブモデリングを提案している。WaveGrad や DiffWave は全ての雑音除去ステップに対して 1 つのモデルを学習しているが、生成ステップの序盤は雑音成分が優勢、生成ステップの終盤は音声成分が優勢と状況は大きく異なる。この点に着目し、雑音スケジュールを分割し（図 9）、それぞれ別々のモデルで学習することにより、合成速度と保ちつつ合成品質を向上できることを示した [16]。

5.5　End-to-end モデル

これまで紹介した波形生成モデルは音響特徴量から音声波形を出力するモデルであるため、TTS においては、別途テキストや音素系列から音響特徴量を推定する音響モデルが必要であった。しかし、音響モデルにより推定される音響特徴量には誤差を含むため、推定した特徴量を用いたファインチューニングを行ったと

しても、多少の音質劣化は避けられない。この問題を解決するために、テキストや音素系列から音声波形を1つのニューラルネットで直接生成可能なEnd-to-endモデルがいくつか提案されている。その中でも、HiFi-GANをデコーダとするVITS [44]は原音に匹敵する非常に高音質なEnd-to-endモデルである。NICTにおいても、更なる高品質化を目指し、End-to-endモデルの検討も行っている。

6　おわりに

これらの深層生成モデルを用いることにより、単一話者TTSモデルにおいては、自然音声と同等の音声波形をCPUのみでリアルタイムで生成できるまでに至っており、NICTでは2022年3月よりVoiceTra®にて日英中韓越の5言語において、CPUのみでリアルタイムに動作する高品質ニューラルTTSを採用しており、更なる多言語化、高品質化に向けて研究開発を行っている。また、音声合成研究をより一層加速させるために、NICTから2022年に日本語（男女各20,000文）及び英語（男女各14,000文）の対話調音声合成用コーパスを公開する。

今後は、学習データには含まれない未知話者に対応した複数話者TTSモデル[45]や歌声合成等のフル帯域合成[20]の高精度化等が課題となる。また、基本周波数[46]や話速[18]を自在に制御可能な波形生成モデルの検討も重要であり、いかにデータの範囲外の基本周波数や話速を外装できるかが課題である。

一方、ここまで合成品質が高くなるとそれを悪用する試みも考えられる。合成音声を安全に利用するために、自然音声と合成音声とを適切に見分ける識別技術[47]の開発も重要な課題として取り組む必要がある。

【参考文献】

1 K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, “Speech synthesis based on hidden Markov models,” Proc. IEEE, vol.101, no.5, pp.1234–1252, May 2013.
2 志賀 芳則, 河井 恒, “多言語音声合成システム,” 情報通信研究機構季報、vol.58, no.3/4, pp.19–25, Sept. 2012.
3 H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” Proc. ICASSP, , pp.7962–7966, May 2013.
4 Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.- J. Qian, H. Meng, and L. Deng, “Deep learning for acoustic modeling in para-metric speech generation: A systematic re- view of existing techniques and future trends,” IEEE Signal Process. Mag., vol.32, no.3, pp.35–52, May 2015.
5 K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, “Model integration for HMM- and DNN-based speech synthesis using product-of-experts framework,” Proc. Interspeech, pp.2288–2292, Sept. 2016.
6 A. van den 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,” Proc. SSW9, p.125, Sept. 2016.
7 J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” Proc. ICASSP, pp.4779– 4783, April 2018.
8 S. H. Mohammadi and A. Kain, “An overview of voice con- version systems,” Speech Commun., vol.88, pp.65–82, April 2017.
9 T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, “Subband WaveNet with overlapped single- sideband filterbanks,” Proc. ASRU, pp.698– 704, Dec. 2017.
10 T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, “An investigation of subband WaveNet vocoder covering entire audible frequency range with limited acoustic features,” Proc. ICASSP, pp.5654–5658, April 2018.
11 T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, “Improving FFTNet vocoder with noise shaping and subband approaches,” Proc. SLT, pp.304–311, Dec. 2018.
12 T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, “Investigations of real-time Gaussian FFTNet and parallel WaveNet neural vocoders with simple acoustic features,” Proc. ICASSP, pp.7020–7024, May 2019.
13 T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, “Real- time neural text-to-speech with sequence-to-sequence acoustic model and WaveGlow or single Gaussian WaveRNN vocoders,” Proc. Interspeech, pp.1308–1312, Sept. 2019.
14 T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, “Tacotron-based acoustic model using phoneme alignment for practical neural text-to-speech systems,” Proc. ASRU, pp.214–221, Dec. 2019.
15 T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, “Transformer-based text-to-speech with weighted forced attention,” Proc. ICASSP, pp.6729–6733, May 2020.
16 T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, “Noise level limited sub-modeling for diffusion probabilistic vocoders,” Proc. ICASSP, pp.6014–6018, June 2021
17 T. Okamoto, T. Toda, and H. Kawai, “Multi-stream HiFi- GAN with data-driven waveform decomposition,” Proc. ASRU, pp.610–617, Dec. 2021.
18 T. Okamoto, K. Matsubara, T. Toda, Y. Shiga, and H. Kawai, “Neural speech-rate conversion with multispeaker WaveNet vocoder,” Speech Commun., vol.138, pp.1–12, March 2022.
19 K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, Y. Shiga, and H. Kawai, “Investigation of training data size for real-time neural vocoders on CPUs,” Acoust. Sci. Tech., vol.42, no.1, pp.65–68, Jan. 2021.
20 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.
21 K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, and H. Kawai, “Comparison of real-time multi-speaker neural vocoders on CPUs,” Acoust. Sci. Tech., vol.43, no.2, pp.121–124, March 2022.
22 岡本 拓磨, “ニューラルネットワークに基づく音声波形生成モデル,” 日本音響学会誌、vol.78, no.6, pp.328–337, June 2022.
23 X. Gonzalvo, S. Tazari, C. an Chan, M. Becker, A. Gutkin, and H. Silen, “Recent advances in google real-time HMM- driven unit selection synthesizer,” Proc. Interspeech, pp.2238–2242, Sept. 2016.
24 H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “Restructuring speech representations using a pitch- adaptive time-frequency 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, April 1999.
25 M. Morise, F. Yokomori, and K. Ozawa, “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, July 2016.
26 A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, “Speaker-dependent WaveNet vocoder,” Proc. Interspeech, pp.1118–1122, Aug. 2017.
27 A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lock- hart, L. C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, “Parallel WaveNet: Fast high-fidelity speech synthesis,” Proc. ICML, pp.3915–3923, July 2018.
28 K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, “An investigation of noise shaping with perceptual weighting for WaveNet-based speech generation,” Proc. ICASSP, pp.5664–5668, April 2018.

29 N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, “Neural speech synthesis with Transformer network,” Proc. AAAI, pp.6706–6713, Jan. 2019.
30 R. Prenger, R. Valle, and B. Catanzaro, “WaveGlow: A flow-based generative network for speech synthesis,” Proc. ICASSP, pp.3617–3621, May 2019.
31 Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “FastSpeech: Fast, robust and controllable text to speech,” Proc. NeurIPS, pp.3165–3174, Dec. 2019.
32 I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. Skerry-Ryan, and Y. Wu, “Parallel Tacotron 2: A non-autoregressive neural TTS model with differentiable duration modeling,” Proc. Interspeech, pp.141–145, Aug. 2021.
33 J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech syn- thesis,” Proc. NeurIPS, pp.17022–17033, Dec. 2020.
34 N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, “Efficient neural audio synthesis,” Proc. ICML, pp.2415–2424, July 2018.
35 J.-M. Valin and J. Skoglund, “LPCNet: Improving neural speech synthesis through linear prediction,” Proc. ICASSP, pp.5826–7830, May 2019.
36 Y.-C. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, “Non-parallel voice conversion system with WaveNet vocoder and collapsed speech suppression,” IEEE Access, vol.8, pp.62094–62106, 2020.
37 D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, “Improved variational inference with inverse autoregressive flow,” Proc. NIPS, pp.4743–4751, Dec. 2016.
38 W. Ping, K. Peng, and J. Chen, “ClariNet: Parallel wave generation in end-to-end text-to-speech,” Proc. ICLR, May 2019.
39 D. Rezende and S. Mohamed, “Variational inference with normalizing flows,” Proc. ICML, pp.1530–1538, July 2015.
40 I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” Proc. NIPS, pp.2672–2680, Dec. 2014.
41 J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” Proc. NeurIPS, Dec. 2020.
42 N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, “WaveGrad: Estimating gradients for waveform generation,” Proc. ICLR, May 2021.
43 Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, “DiffWave: A versatile diffusion model for audio synthesis,” Proc. ICLR, May 2021.
44 J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” Proc. ICML, pp.5530–5540, July 2021.
45 E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, “YourTTS: Towards zero-shot multi- speaker TTS and zero-shot voice conversion for everyone,” Proc. ICML, pp.2709–2720, July 2022.
46 Y.-C. Wu, T. Hayashi, T. Okamoto, H. Kawai, and T. Toda, “Quasi-Periodic Parallel WaveGAN: A non-autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol.29, pp.792–806, 2021.
47 俵 直弘, “話者認識システムとなりすまし対策,” 日本音響学会誌、vol.78, no.6, pp.338–346, June 2022.

岡本 拓磨 （おかもと たくま）

ユニバーサルコミュニケーション研究所
先進的音声翻訳研究開発推進センター
先進的音声技術研究室
主任研究員
博士（情報科学）
音場制御、音声合成

【受賞歴】
2022 年　日本音響学会 第 9 回学会活動貢献賞
2018 年　日本音響学会 第 57 回佐藤論文賞
2012 年　日本音響学会 第 32 回粟屋潔学術奨励賞