5R-03

# 高品質なリアルタイム音声生成のための自己回帰型ニューラルポストフィルタの検討

長沢 一生 †　　菅野 由弘 ‡
早稲田大学 †　　早稲田大学 ‡

## 1 はじめに

深層学習技術の活用により，テキスト音声合成や音声変換などの音声生成システムの性能は近年大きく向上した．しかし，学習データが少ない場合や，複数話者の音声で学習を行う場合など，条件によっては依然として高品質な音声の生成は難しい．また，音声生成システムの運用において，品質向上のため，音声特徴量を生成するモデルの学習をやり直すことがあるが，これにより生成される音声のキャラクタ性が意図せず変化してしまう場合がある．本稿では，既存の音声生成システムの生成する音声に対し，自己回帰型のニューラルネットワークにより時間領域上で補正を行うことで品質を向上させることを検討する．

## 2 関連研究

### 2.1 音声強調

音声強調の分野では，ノイズや残響を含んだ音声から歪みの少ない音声を抽出する様々な手法が提案されている．これらの手法は，その手法が仮定している条件 (目標音声とノイズは独立，一つの入力音声に対応する目標音声はただ一つなど) の下では非常によく動作する．しかし，テキスト音声合成や音声変換において生成された合成音声から自然な音声へ補正を行う場合においては，この仮定が満たされず，単純に手法を適用するのは難しい．

WaveCycleGAN2 [1] は，合成音声と自然な音声を相互に変換する循環型モデルを用いることにより，入力音声と目標音声の一対一対応を実現し，品質の向上を可能にしている．この手法は単一話者データセットにおいて合成音声の大幅な高品質化に成功しているが，GAN の学習は一般的に難易度が高く，また，WaveCycleGAN2 は多くの種類の損失関数を使用するため，学習を更に難しくしている．

### 2.2 ニューラルボコーダ

ニューラルボコーダは，ニューラルネットワークを用いて音響特徴量から音声を合成するモデルであり，近年の音声生成技術向上の根幹となった技術である．比較的高速に動作するニューラルボコーダのうち，自己回帰モデルを使用したものとして WaveRNN [2] が挙げられる．WaveRNN は，1 層の GRU と続く 2 層の全結合層によって，高品質な音声波形を高速に生成する．GRU の行列をスパース化した WaveRNN は，モバイル CPU 上で実時間での動作を達成している．多人数・多言語での高品質な波形生成を達成した RNN_MS [3] も，WaveRNN を元にしたシステムである．

## 3 提案法

WaveRNN を用いて，合成音声の品質を向上させる手法を提案する．提案法は，学習に用いるパラレルデータの生成方法と，WaveRNN の学習方法の 2 部からなる．

### 3.1 パラレルデータの生成

入力として x-vector [4] を与え条件付けを行うことで変換先話者を指定できる，差分スペクトル補正 [5] に基づいた音声変換システムを構築する．このシステムにより，自然音声に対しランダムな話者の x-vector を使用して変換を行い，さらに元の自然音声から抽出した x-vector を使用して再度変換を行うことで，合成音声を模した品質の低い音声を生成する．この方法により生成された音声は，元音声と発話内容および基本周波数 $F_0$ は常に等しいが，波形の形状は全く異なる場合もある．

### 3.2 WaveRNN の学習

WaveRNN は，前述の低品質な音声波形とそのメルスペクトログラムを入力として，対応する自然音声との残差の確率分布および自然音声のメルスペクトログラムの予測値を出力する．損失関数として，予測分布の負の対数尤度に加えて，メルスペクトログラムの平均絶対誤差を併せて使用する．通常の WaveRNN との最も大きい差は入力として音声波形が与えられることだが，この音声は目標音声と全く

Autoregressive Neural Post-Filters for High-Quality Real-Time Speech Generation
† Issei Nagasawa, Waseda University
‡ Yoshihiro Kanno, Waseda University

表 1　入力音声・自然音声それぞれとの対数 $F_0$ の平均絶対誤差

| | 入力音声 [semitone] | 自然音声 [semitone] |
|---|---|---|
| 入力音声 | 0 | 0.465 |
| 提案法 | 0.467 | 0.282 |
| 提案法 (波形入力無し) | 0.667 | 0.517 |

等しい周期の波形を持つことから，モデルは入力音声に正確に等しい周波数の波形を生成するように学習が進むと期待される．

## 4　実験的評価

### 4.1　学習条件

パラレルデータ生成のための音声変換モデルの学習には，JVS コーパス [6] のうち 80 話者の音声を使用した．WaveRNN の学習には，自然音声として青空朗読 [7] より約 100 時間の朗読音声を使用した．音声のサンプリング周波数は 16kHz である．

### 4.2　客観評価: $F_0$ の一致性

入力音声のキャラクタ性が高品質化された出力音声でも維持されていることを示す指標のひとつとして，出力音声の，入力音声および自然音声との対数 $F_0$ の平均絶対誤差を計算した．ただし，大きな $F_0$ のずれは，モデルの設計に由来する性質よりも，モデルの学習度合いや推論の不安定さ，あるいは $F_0$ 分析エラーに起因する面が大きいと考えられるため，いずれかの合成音声または入力音声に自然音声と 6 半音以上のずれがあった箇所は計算から除外した．検証には，WaveRNN の学習に使用したデータに含まれる 9 名の話者の，学習に使用していない音声を用いた．比較のため，メルスペクトログラムのみを入力として与え，残差でなく目標の波形を直接予測する WaveRNN についても学習を行い，同様に誤差の計算を行った．表 1 に示した結果から，波形を入力として与える WaveRNN は，与えないものと比較し，入力した音声にも目標となる自然音声にも近い $F_0$ 軌跡の音声を出力することが示された．

### 4.3　主観評価: 自然性

音質の自然性について 5 段階 MOS 評価を行った．検証には，青空朗読から既知の 9 話者，JVS コーパスから未知の 10 話者，それぞれ 1 文の音声を使用した．実験参加者は 16 名であり，各参加者は自然音声，入力音声，提案モデルの出力音声のそれぞれを評価した．既知話者と未知話者それぞれについての結果を図 1 に示す．提案法により，いずれの話者グループにおいても自然性の改善が見られたが，出力音声の品質は自然音声に遠く及ばなかった．

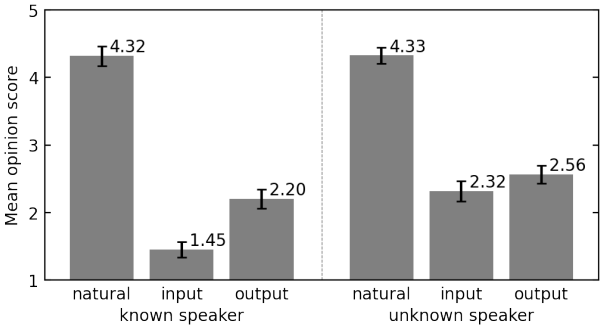


図 1　音声の自然性に関する MOS 評価結果．エラーバーは 95% 信頼区間を示す．

## 5　まとめ

自己回帰モデルを用い，合成音声の品質をキャラクタ性を維持したまま時間領域で直接向上させる方法の検討を行った．結果として，合成音声を模したデータにおいて，提案法により未知話者と既知話者のいずれの音声でも自然性が向上することが確認できたが，十分と言える品質には到達しなかった．品質の更なる向上と，テキスト音声合成や音声変換システムへの実際の適用が今後の課題である．

## 参考文献

[1] Kou Tanaka et al. WaveCycleGAN2: Time-domain Neural Post-filter for Speech Waveform Generation. *arXiv preprint arXiv:1904.02892*, 2019.

[2] Nal Kalchbrenner et al. Efficient Neural Audio Synthesis. In *Proc. ICML*, pages 2410–2419. PMLR, 2018.

[3] Jaime Lorenzo-Trueba et al. Towards achieving robust universal neural vocoding. In *Proc. Interspeech*, pages 181–185, 2019.

[4] David Snyder et al. X-vectors: Robust dnn embeddings for speaker recognition. In *Proc. ICASSP*, pages 5329–5333. IEEE, 2018.

[5] Kazuhiro Kobayashi et al. Statistical singing voice conversion with direct waveform modification based on the spectrum differential. In *Proc. Interspeech*, pages 2514–2518, 2014.

[6] Shinnosuke Takamichi et al. JVS corpus: free Japanese multi-speaker voice corpus. *arXiv preprint arXiv:1908.06248*, 2019.

[7] 一般社団法人 青空朗読. 青空朗読. `https://aozoraroudoku.jp/`.