WaveNet を用いた歌唱表現が可能な歌声合成方式の検討

Study of Singing Voice Synthesis generate song expression Using WaveNet

金子 隼人

Hayato Kaneko

岡山大学 阿部研究室

Abe Laboratory, Okayama University

概要 本研究では，WaveNet を用いた歌唱表現付きの歌声合成方式を提案する．歌唱表現は，ビブラートやオーバーシュートといった音高変動を含み，歌声を特徴づけるうえで重要な要素である．本研究では，WaveNet の学習時に補助特徴量として音符の音高のような楽譜情報を使用することにより，歌唱表現付きの自然な歌声を合成することを目指す．

1 はじめに

歌声合成とは，任意の楽曲の歌声を人工的に作り出す技術であり，個人の歌声の再現 [1] や病気で声を失った患者の歌声を再現する手法 [2] など，障がい者支援の観点からの利用も考えられている．歌声合成に似た技術として音声合成という技術がある．音声合成は，人間の声を人工的に作り出す技術である．音声合成では，音声のアクセントやイントネーションを再現するために，音高の時間的変化の軌跡の概形を再現する必要がある．しかし，歌声は話し声と比較して発声の高さや強さの変動幅が広く，より複雑な特性を持つことが知られており，特に歌声の音高の時間的変化の軌跡には，歌声固有の特徴が表れ，歌声合成ではこの特徴も再現する必要がある．

近年，VOCALOID[3] のような波形接続型の歌声合成とは異なるアプローチとして，Deep Neural Network(DNN) を用いた歌声合成方式が検討されている．合成歌声の品質向上の試み [4] だけでなく，ユーザの歌い方を真似することが可能な歌声合成システム [5] も提案されており，歌声合成の多様化が進んでいる．

本研究では，WaveNet[6] を用いた歌唱表現付きの歌声合成の方式を提案する．歌唱表現は，ビブラートやポルタメントなどの音高変動，音量の変化や声質が含まれ，歌声を特徴づけるうえで重要な要素である．近年提案されている DNN を用いた歌声合成の手法は 2 段階に分かれており，特徴量を生成する特徴量生成部と特徴量から歌声を合成する音声信号生成部に分かれている．楽譜情報を使った歌声合成の研究では，文献 [7] のように特徴量生成部で楽譜情報を使用する方式が一般的である．提案方式では音声信号生成部である WaveNet の補助特徴量として楽譜情報を使用することにより，歌唱表現付きの歌声を合成する．歌唱表現は歌手が任意につける表現や無意識につける表現があり，歌手は楽譜を見て次に歌う音高などを意識しながら歌っているため，楽譜情報により歌声に表れる歌唱表現を予測できると期待している．

図 1: 提案方式の概要図

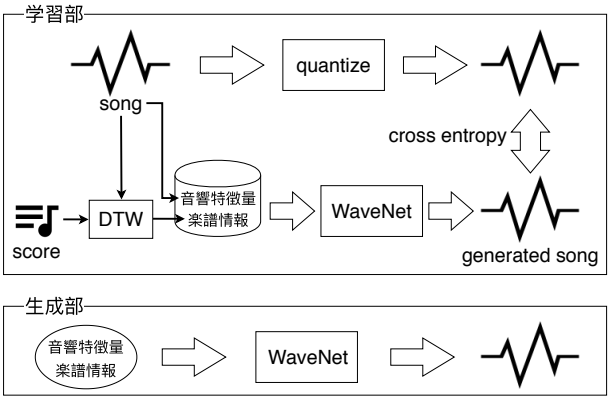

図 2: 音声信号生成部の概要図

2 WaveNet

WaveNet は，過去の波形から直接未来の波形を予測する Convolutional Neural Network である．WaveNet は，複数の Residual block から構成されており，各 Residual block 中で dilated causal convolution を一回おこなう．WaveNet では，補助特徴量を追加することにより，生成される波形を制御することが可能である．

3 提案方式

本稿では，WaveNet を用いた任意の楽譜から歌唱表現付きの歌声合成方式を検討する．提案方式の概要を図 1 に示す．

提案方式は特徴量生成部と音声信号生成部から成り，特徴量生成部では楽譜情報から音響特徴量を生成し，音声信号生成部では音響特徴量と楽譜情報から歌声を合成する．楽譜情報とは楽譜から得られる情報のことであり，各音符の音高や継続長などが含まれる．

学習時には楽譜からそれに対応する歌声を生成するように学習を行う．特徴量生成部では楽譜情報から対

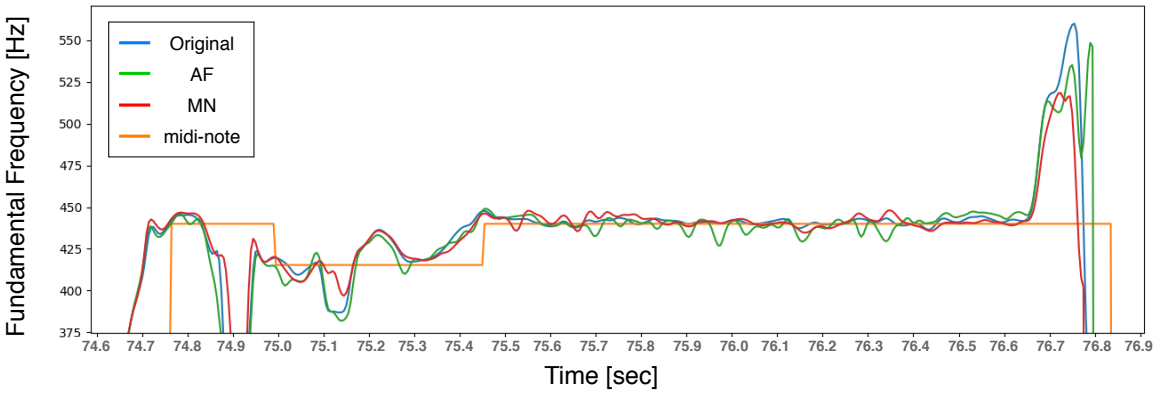

図 3: 基本周波数の時間的変動

応する歌声の音響特徴量を再現するように学習し，音声信号生成部では音響特徴量から音声波形を再現するように学習する．合成時には任意の楽譜情報を入力することで，その楽譜を学習した歌手が歌った場合の歌声を再現することを目指す．

図 2 に示すように WaveNet を用いた音声信号生成部は，学習部と生成部に分かれている．学習部では歌声から得られた音響特徴量とそれに対応する楽譜情報を補助特徴量として用いて，元データの歌声を学習する．生成部では，音響特徴量と楽譜情報を用いて歌唱表現付きの歌声の合成を行う．

4 評価実験

提案方式では，WaveNet の補助特徴量に楽譜情報を追加することにより，音声に歌唱表現を付与することを目指している．しかし，楽譜情報はステップ状の情報であり，歌唱表現は非線形に変化するものであるため，楽譜情報から歌唱表現を予測できるかは不明である．

このことから，ステップ状の楽譜情報から非線形の歌唱表現を再現できるかを確認する実験を行う．

4.1 実験条件

学習データとして，東北きりたん歌唱データベース [8] を用いた．このデータベースは，50 曲の歌声の音声ファイル (約 30 分) とそれに対応する楽譜で構成されており，女性歌手 1 名が J-pop の曲を歌った歌声の音声と，歌声に合わせて人手で調整がされた楽譜ファイルである．iteration 数は 200,000 とし，ミニバッチサイズは 1，最適化手法には Adam[9] を用いた．Residual block の数は 30 とし，dilation は $[2^0, 2^1, 2^2, ..., 2^9]$ を 3 回繰り返した．補助特徴量として，音声をスペクトログラム分析することによって抽出された 80 次元スペクトログラムと，楽譜から抽出された音符の音高を 48 次元 one-hot-vector に変換したものを用いた．

表 1: 楽譜の音高との比較

	log-F0-RMSE[cent]
AF	18.63
MN	18.27
元データ	18.97

4.2 評価実験用音声

客観評価実験のために，東北きりたん歌唱データベースから学習に使用していない 5 曲を使用した．音響特徴量としてスペクトログラムを，楽譜情報として楽譜の音高を WaveNet の補助特徴量として用いた．スペクトログラムのみを補助特徴量に用いたモデルと，スペクトログラムと楽譜情報を補助特徴量に用いたモデルを学習し，それぞれのモデルで歌声を合成し使用した．以下，スペクトログラムのみを補助特徴量に用いたモデルを AF，スペクトログラムと楽譜情報を補助特徴量に用いたモデルを MN と呼ぶ．

4.3 客観評価実験

ステップ状の楽譜情報を補助特徴量として使用しても，非線形の歌唱表現が再現されるかを確認するために客観評価実験を行う．基本周波数の時間的変動を確認するとともに，客観評価指標として対数基本周波数の平均二乗誤差 (log-F0-RMSE) を算出する．基本周波数は WORLD[10] により抽出された．基本周波数とは音の高さに対応する音響特徴量であり，時間的変動を確認することにより一部の歌唱表現を確認することができる．また，人間の聴覚特性は対数的であり，聴覚上では一定の間隔で音程が上がっているように聞こえても，基本周波数は指数的に増加している．このことから，対数基本周波数を評価指標として使用する．

4.4 実験結果

各音声の基本周波数の時間的変動と楽譜の音高の変化を図 3 に示す．76.7 sec 付近ではプレパレーションという歌唱表現が見られる．これは音高が変化する直前にその変化とは逆方向に音高が変化する特徴である．AF と MN の歌声においても同様に表れていることから，AF と MN のどちらもこの特徴を正しく学習し，歌手の特徴を再現できている．しかし，MN は AF や元データと比べ楽譜の音高との差が小さくなっている．MN では楽譜の音高を補助特徴量に使用したことによる影響と考えられる．このように MN が元データより楽譜の音高に近づいているという特徴は，例えば 75.1 sec や 76.0 sec 付近で確認できる．

また，表 1 に楽譜の音高と各音声の音高との log-F0-RMSE を示す．値は生成された歌声と元データの異なり度合を距離尺度で表しており，全フレームにわたり計算されている．この比較では MN が最も低い値を示した．このことからも，MN は楽譜の音高との差が小さくなっていることがわかる．

しかし，楽譜の音高との差が最も大きいのは元データであり，WaveNet で合成した歌声はその値よりも小さくなっていることから，WaveNet で完璧に特徴を再現できているわけではなく，元データより特徴が弱い歌声が生成されていると考えられる．したがって，フレーム単位の音符の音高を使用するだけでは特徴を再現することが困難であると考えられる．歌手は楽譜を見ながら次に歌う音符の音高や継続長を意識しながら歌っており，歌唱表現も後の楽譜情報に影響を受けていると考えられる．これらのことから，前後の楽譜情報を補助特徴量に使用するべきだと考えられる．

5 まとめ

本稿では，WaveNet を用いた歌唱表現が可能な歌声合成方式を提案した．WaveNet の学習時に補助特徴量として楽譜情報を使用することにより，歌唱表現付きの歌声を合成した．生成された歌声の歌唱表現の再現性を基本周波数の時間的変動を観察することで確認した．また，ステップ状の特徴量を入力しても，非線形な歌声の特徴を再現できることを確認した．

今後の課題として，本論文で扱わなかった歌唱表現を再現するための方法を検討することや，前後の楽譜情報の使用が挙げられる．

参考文献

[1] M. Blaauw, J. Bonada, R. Daido, “Data efficient voice cloning for neural singing synthesis,” Proceedings of ICASSP 2019, 2019.

[2] S. Arik *et al.*, “Neural voice cloning with a few samples,” Proceedings of Advances in Neural Information Processing Systems, 2018.

[3] 剣持秀紀，大下隼人，“歌声合成システム VOCALOID-現状と課題，” 情報処理学会研究報告, 音楽情報科学 (MUS)， 2008，2008.

[4] M. Nishimura *et al.*, “Singing Voice Synthesis Based on Deep Neural Networks.,” Proceedings of Interspeech, 2016.

[5] T. Nakano, M. Goto, “VocaListener: A singing-to-singing synthesis system based on iterative parameter estimation,” Proceedings of Sound and Music Computing Conference, 2009.

[6] A. Oord *et al.*, “WaveNet: A generative model for raw audio,” arXiv preprint arXiv:1609.03499, 2016.

[7] K. Nakamura *et al.*, “Fast and High-Quality Singing Voice Synthesis System based on Convolutional Neural Networks,” arXiv preprint arXiv:1910.11690, 2019.

[8] “東北きりたん歌唱データベース，” `https://zunko.jp/kiridev/login.php`Accessed: 2020.01.27.

[9] D. P. Kingma, J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[10] M. Morise, F. Yokomori, K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE TRANSACTIONS on Information and Systems, 99, 2016.