3N-05

WaveNetを用いた音符系列に対する歌唱F0軌跡の生成

和田 雄介　　糸山 克寿　　吉井 和佳
京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

歌声合成は，コンピュータに任意の楽曲を歌わせることを可能にする技術である．ある歌唱に対して，声質変換および歌唱表現転写を適用することで，別の歌唱に作り変えることができる．歌唱表現の転写は，ある歌手の楽曲の歌声音符系列の推定，その歌手の歌い方の学習，別の楽曲から推定された歌声音符系列への表現付与の3段階で行う．付与すべき歌唱表現のうち，歌唱F0軌跡は，歌唱の音高に加えて，ビブラートなどの歌唱表現の特徴を含んでおり，自然かつ表情豊かな歌声を合成するために重要である．

これまでに，歌声に対する様々な声質変換手法 [1, 2] や音符系列推定手法 [3, 4] が提案されている．また，音符系列から歌声を合成する手法には，音符系列に対する変動成分のパラメータを人手で与える VOCALOID [5] がある．音符系列から歌唱F0軌跡を生成する手法として，大石ら [6] は，楽譜成分に対する変動を表現する自己回帰関数を用いて，歌唱F0軌跡をモデル化した．この手法では，線形な自己回帰関数を仮定するのに対して，近年，深層学習を用いて非線形な自己回帰モデルを形成する手法が提案されている [7, 8]．SampleRNN [7] は再帰型ニューラルネットワークを用いるのに対して，WaveNet [8] は畳み込みニューラルネットワークを用いる．このうち WaveNet は，畳み込み層の積み重ねによって広範囲の情報を直接考慮できる．

本稿では，WaveNet を用いて歌唱F0軌跡をモデル化し生成する手法を提案する．深層自己回帰モデルの構築には，再帰型ニューラルネットワークや畳み込みニューラルネットワークの使用が考えられるが，再帰型ニューラルネットワークの学習には，大量の学習データが必要である．本手法では，少ない学習データに対しても頑健に動作するモデルを構築するため，畳み込みニューラルネットワークをベースとし，高い表現力を持つ WaveNet を使用する．

2. 提案手法

本章では，音符系列を入力として，歌声として自然な周波数方向の変動が付与された歌唱F0軌跡を WaveNet を用いて生成する手法について説明する．本手法の入力音符系列は，フレーム単位の対数周波数 (単位は cent) の系列 $\boldsymbol{h} = \{h_t\}_{t=1}^{T}$ とする．ただし，t はフレーム番号，T は系列の個数である．出力の歌唱F0軌跡は，対数周波数 (単位は cent) の系列 $\boldsymbol{y} = \{\boldsymbol{y}_t\}_{t=1}^{T}$ である．$\boldsymbol{y}_t$ は，one-hot vector で表すとする．本手法では，$\boldsymbol{h}$ を補助特徴量とした，$\boldsymbol{y}$ に対する WaveNet を学習し，それを用いて新たな音符系列 $\boldsymbol{h}'$ に対する歌唱F0軌跡を生成する．

WaveNet により歌唱F0軌跡が逐次推定される過程を，図1に示す．WaveNet は，歌唱F0軌跡 $\boldsymbol{y}$ の同時確率

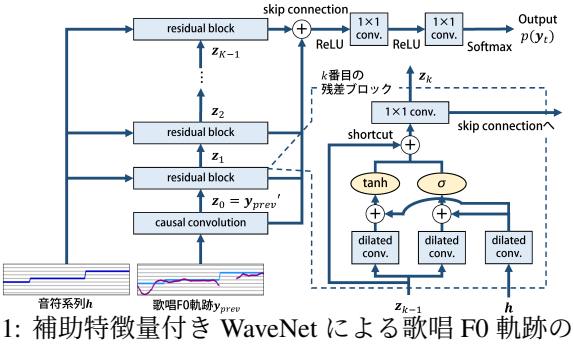


図 1: 補助特徴量付き WaveNet による歌唱 F0 軌跡の逐次予測．

$$p(\boldsymbol{y}) = \prod_{t=1}^{T} p(y_t | y_1, y_2, \cdots, y_{t-1}) \tag{1}$$

を表現する．式 (1) では，過去の全てのサンプルを考慮しているが，ネットワークの大きさは有限であり，実際に考慮できるサンプルの数には限りがある．そのため，WaveNet は，式 (1) の同時確率を，

$$p(\boldsymbol{y}) \approx \prod_{t=1}^{T} p(y_t | y_{t-R}, y_{t-R+1}, \cdots, y_{t-1}) \tag{2}$$

によって近似する．式 (2) 中の R は，考慮できる過去のサンプル数であり，受容野と呼ばれる．式 (2) の同時確率は，dilated convolution (DC) と呼ばれる穴開きのフィルタを用いた畳み込み層，活性化関数，shortcut を含む残差ブロックの積み重ねで表現される．受容野 R は，全ての DC におけるフィルタの穴開きの数の合計となる．

系列 $\boldsymbol{y}_{\text{prev}} = \{y_{t-R}, y_{t-R+1}, \cdots, y_{t-1}\}$ から，$\boldsymbol{y}_t$ を予測する過程について説明する．$\boldsymbol{y}_{\text{prev}}$ は，まず 1×1 の (フィルタサイズ 1 の) 畳み込み層を経由して $\boldsymbol{y}'_{\text{prev}}$ に変換されたのち，最初の残差ブロックに入力される．k 番目の残差ブロックの出力 $\boldsymbol{z}_k$ は，

$$\boldsymbol{z}_k = \tanh(W_{f,k} * \boldsymbol{z}_{k-1}) \odot \sigma(W_{g,k} * \boldsymbol{z}_{k-1}) \tag{3}$$

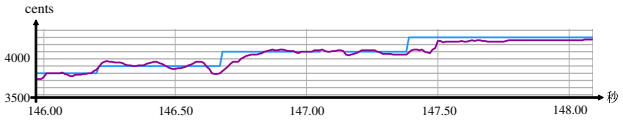
のように，直前の残差ブロックの出力 $\boldsymbol{z}_{k-1}$ に対する DC 層の出力と，活性化関数によって決まる．ただし，$\boldsymbol{z}_0 = \boldsymbol{y}'_{\text{prev}}$ であり，$\odot$ は要素積，$\sigma(\cdot)$ はシグモイド関数，$W_{f,k}, W_{g,k}$ はそれぞれ DC のフィルタ，$*$ は畳み込み演算を表す．全ての残差ブロックの出力は skip connection によって統合され，最終的なネットワークの出力は，softmax 関数による $\boldsymbol{y}_t$ の各要素の生起確率となる．

WaveNet は，補助特徴量 $\boldsymbol{h} = \{h_1, h_2, \cdots, h_{T'}\}$ を用いた条件付き確率 $p(\boldsymbol{y}|\boldsymbol{h})$ をモデリングできる．このとき，式 (1) の同時確率は，

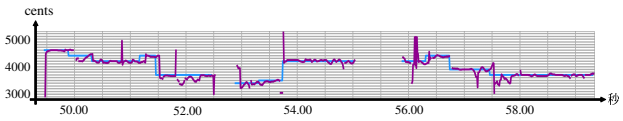
$$p(\boldsymbol{y}|\boldsymbol{h}) = \prod_{t=1}^{T} p(y_t | y_1, y_2, \cdots, y_{t-1}, \boldsymbol{h}) \tag{4}$$

となる．本手法では，補助特徴量として，フレーム単位の音符系列を使用する．これにより，WaveNet による歌唱F0軌跡の学習及び生成に対して，音符情報を条件付ける．具体的には，式 (3) の計算を，

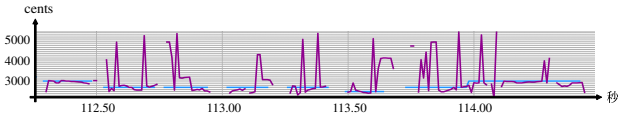
Generation of Singing F0 Trajectories for Musical Note Sequences Based on WaveNet: *Yusuke Wada, Katsutoshi Itoyama, and Kazuyoshi Yoshii (Kyoto Univ.)*



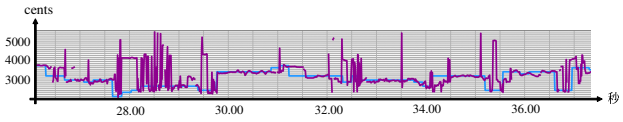
(a) 約 2 秒の入力音符系列に対して，その付近に生成された歌唱 F0 軌跡.



(b) 約 10 秒の入力音符系列に対して，その付近に生成された歌唱 F0 軌跡.



(c) 約 2 秒の入力音符系列に対して，そこから大きく外れた箇所に生成された歌唱 F0 軌跡.



(d) 約 10 秒の入力音符系列に対して，そこから大きく外れた箇所に生成された歌唱 F0 軌跡.

図 2: 生成された歌唱 F0 軌跡の例．青色の線が入力された音符系列を，紫色の線が出力された歌唱 F0 軌跡を表す．灰色の線は，縦軸は 100cent 間隔の，横軸は 0.5 秒間隔のグリッドである．

$$\boldsymbol{z}_k = \tanh(W_{f,k} * \boldsymbol{z}_{k-1} + W'_{f,k} * \boldsymbol{h}) \odot \sigma(W_{g,k} * \boldsymbol{z}_{k-1} + W'_{g,k} * \boldsymbol{h}) \quad (5)$$

のように修正する．ただし，$W'_{f,k}, W'_{g,k}$ は，それぞれ $W_{f,k}, W_{g,k}$ と同じ大きさの DC のフィルタを表す．

歌唱 F0 軌跡の生成は，学習済みモデルから歌唱 F0 を所望の回数サンプリングすることで行う．WaveNet は自己回帰モデルであるため，過去に生成されたサンプルを新たなサンプルの予測に用いる．

3. 評価実験

提案手法による歌唱 F0 軌跡の生成例を示す．RWC 研究用音楽データベース [9] のポピュラー楽曲 100 曲のうち，63 曲をモデルの学習に用いた．音符系列および歌唱 F0 軌跡は，データベース内のアノテーションデータ [10] を用いた．これらのアノテーションデータは，10 ミリ秒間隔の対数周波数 (連続値) の系列である．このうち，C2 (1500cent)–C6 (6300cent) の範囲にない値は無音とみなして 10cent 単位にクオンタイズしたのち，481 次元 + 無音を表すラベルの 482 次元の one-hot vector に変換したものを用いた．

WaveNet の構成について，残差ブロックの個数は 24 個とし，各ブロック内の DC の穴開きは，入力に近い層から順に $1, 2, 4, \cdots, 128$ を 3 回繰り返したものとした．また，残差ブロックの前と内部にある 1×1 畳み込みのチャネル数は 256 とし，skip connection から出力の間にある 1×1 畳み込みのチャネル数は 1024 とした．パラメータの更新は，30 素片を 1 ミニバッチとして，ハイパーパラメータ $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999$ の Adam[11] によって行った．

図 2 に，約 2 秒間の系列と約 10 秒間の音符系列 2 つずつに対する歌唱 F0 軌跡の生成結果の例を，入力音符系列とともに示す．入力には，データベース内の学習に用いていない 37 曲分の音符系列を用いた．(2a) では，入力楽譜系列の付近にほぼ正確に歌唱 F0 軌跡が出力された．また，(2b) のように長い系列でも，歌い始めなどの箇所で入力から外れた値が出力されているものの，概ね入力の付近に出力されている．一方．(2c) や (2d) のように，F0 の値が入力から大きく外れた場所に出力されてしまう例もあった．(2c) および (2d) のどちらも，同じ音高が続く区間で入力から大きく外れた値が出力される傾向にあり，これは特に (2d) の 28–30 秒において顕著である．

このような例が出力されるのは，WaveNet の学習および生成において，未来の音符系列の情報が考慮されていないことが原因であると考えられる．例えば (2c)(2d) における外れ値は，同じ音高の音符が続くという情報によって抑制できる．また，歌唱表現は，過去と現在の系列のみならず，後続する音形にも依存する．例えば，ビブラートは，今歌っている音高がそれ以降も長く続くときに現れやすく，オーバーシュートは，休符が終わり歌い始める箇所に表れやすい．そのため，現状の補助特徴量に加えて，数フレーム先の音符系列を入力することで，より自然な歌唱 F0 軌跡を出力できると考えられる．

4. おわりに

本稿では，WaveNet を用いて，入力された音符系列に対して歌唱 F0 軌跡を生成する手法を提案した．実験の結果，ある程度自然な歌唱 F0 軌跡が生成されることが確認された．今後は，入力補助特徴量に数フレーム先の音符系列を追加して学習し，それが生成結果に与える影響を調査する予定である．本手法の応用として，特定の歌唱者の歌唱 F0 軌跡のみを用いてこのモデルを学習することにより，その歌唱者の歌唱表現の特徴を他の歌唱に転写するということが考えられる．また，音量軌跡などの歌唱表現も本手法と同様にモデル化できると考えられ，さらに高精度な転写が期待できる．

謝辞 本研究の一部は、JSPS 科研費 26700020, 16H01744 および JST ACCEL No. JPMJAC1602 の支援を受けた．

参考文献

[1] F. Villavicencio and J. Bonada. Applying Voice Conversion To Concatenative Singing-Voice Synthesis. *Proc. Interspeech*, 2162–2165, 2010.

[2] K. Kobayashi *et al.* Statistical singing voice conversion based on direct waveform modification with global variance. *Proc. Interspeech*, 2754–2758, 2015.

[3] M. P. Ryynänen and A. P. Klapuri. Automatic Transcription of Melody, Bass Line, and Chords in Polyphonic Music. *Computer Music Journal*, 32:72–86, 2008.

[4] L. Yang *et al.* Probabilistic Transcription of Sung Melody Using A Pitch Dynamic Model. *Proc. ICASSP*, 301–305, 2017.

[5] H Kenmochi and H Ohshita. VOCALOID-commercial singing synthesizer based on sample concatenation. *Proc. Interspeech*, 4009–4010, 2007.

[6] 大石 康智 ほか. 畳み込み HMM に基づく歌声の基本周波数制御モデルの提案とそのパラメータ学習方法. **情報処理学会研究報告音楽情報科学** *(MUS)*, 89–96, 2008.

[7] S. Mehri *et al.* SampleRNN: An Unconditional End-to-End Neural Audio Generation Model. *Proc. ICLR*, 1–11, 2016.

[8] A. van den Oord *et al.* WaveNet: A Generative Model for Raw Audio. *arXiv preprint arXiv:1609.03499*, 1–15, 2016.

[9] M. Goto *et al.* RWC Music Database. *Proc. ISMIR*, number October, 229–230, 2003.

[10] M. Goto. AIST annotation for RWC music database. *Proc. ISMIR*, 359–360, 2006.

[11] D. P. Kingma and J. L. Ba. Adam: A Method for Stochastic Optimization, 2014.