1T-01

波形接続型音声合成における歌唱のピッチ変化予測LSTMモデル

田中 瑞穂[1]　平田 圭二[1]　竹川 佳成[1]

1. はじめに

近年，歌声合成システムを使って楽曲を制作するユーザーが増加している．このような歌声を生成するための音声合成ソフトは，表現力を高めるために歌声を調節することができる．しかし，人間歌唱のような物理的な制約がない合成歌唱は，ピッチや声量を無制限に変化させることができる．そのため，ソフト側でパラメータを設定することでポルタメントを一貫して制御している．一方で，制御が可能なパラメータが多いため，初心者は設定に時間がかかる．この問題を解決するため，機械学習によって人間歌唱を真似た合成歌唱生成する研究が注目されている．しかし，合成音声歌唱には，高速歌唱や急激な音程の変化，短いヴィブラートなど，人間では困難な合成音声特有の歌唱技術が多数存在している．これらの歌唱技術は，人間歌唱を用いる従来の手法では学習することができない．

そこで，本研究では，波形接続型合成音声を用いた合成音声ソフトで使用されているテキストデータを元にコーパスの作成を行い，未知の楽譜のパラメータを予測するモデルを提案する．この手法は，合成音声特有の歌唱技術のデータが含まれているため，人間歌唱では困難な歌唱技術にも対応が可能である．また，統計的合成音声ではなく波形接続型合成音声を用いた合成音声ソフトを選択した理由として，学習するパラメータが記号ベースであるため学習時間の短縮につながることが挙げられる．なお，本研究では，抑揚を表現するパラメータに焦点を当てて学習する．

2. 合成音声ソフト UTAU

本研究では，合成音声ソフト UTAU [1] で使用される楽譜 UTAU Sequence Text(以下 UST とする) から抽出した音価，音高，歌詞，ポルタメントのデータをコーパスとする．

[1] 公立はこだて未来大学

一般的にポルタメントは，音から音へ移動する際に，音程を滑らかに変化させる演奏及び歌唱技法を示す．一方，UTAU で用いるポルタメントは，1 つの音の途中でも音程をずらすことができる．UTAU では，ポルタメント機能を用いることで図 1 のように点と点を繋ぐピッチ線を生成し，このピッチ線の形状に従って歌唱中の抑揚が変化する．なお，図 1 の赤い四角はピッチポイントと呼ばれる．UTAU のデフォルトのポルタメントは，2 つのピッチポイントだけでスムーズに音を繋ぐものである．しかし，ピッチポイントの数を増やすことで，ピッチ線を曲げて細かく音程を変化させることが可能となる．

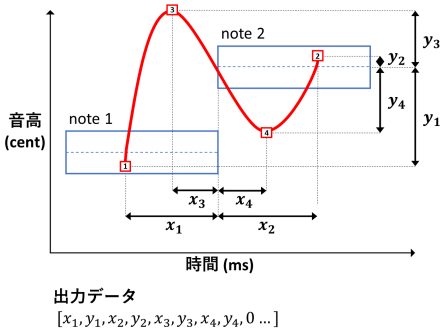

図 1 UTAU におけるポルタメントと出力例

3. パラメータ予測 LSTM モデル

本研究では，機械学習を用いてポルタメントに用いるパラメータを予測する．使用する機械学習モデルは，時系列を考慮しており音の前後関係が学習可能な LSTM [2] である．

入力データは，音価，音高，音のタイプ，前後の音高差を示す 5 つの One-hot ベクトルである．出力データは，ピッチポイントの個数を示す One-hot ベクトルとピッチポイントの座標を示すベクトルである．ピッチポイントの個数は，指定した音に対してヴィブラート及びポルタメントが

出現するか判別するために用いる．

提案モデルは，大きく 2 つに分かれる．1 つは，図 2 のように生成を判別するピッチポイントの個数のパラメータを他のパラメータとまとめて出力するモデルである．もう 1 つは，生成を判別するパラメータを別途学習するモデルである．これは，生成を判別するパラメータを一度学習してから入力に加えた方がまとめて学習するよりも精度が上がると考えたためである．

モデルの学習時は，1 音ずつずらして楽譜全体を学習する．なお，本研究では，前の 31 音を関連付ける 32 次元の LSTM モデルと前の 16 音と後ろの 16 音を関連付ける 33 次元の LSTM モデルをそれぞれ用意している．これは，前の音のみを関連付けた時と前後両方の音を関連付けた時のどちらが精度が上がるか検証するためである．

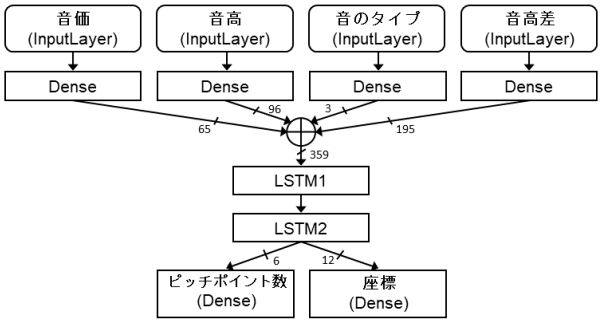

図 2 予測 LSTM モデルのネットワーク

4. ポルタメント再現実験

本実験では，モデルがユーザの作成するパラメータと近い値を予測可能か検証する．モデルは，表 1 の 4 パターンを用いる．縦軸は関連付ける範囲，横軸は生成を判別するパラメータの扱いを示している．

表 1 モデルの種類

	まとめて学習	別途学習
前 31 音	モデル 1	モデル 2
前 16 音及び後 16 音	モデル 3	モデル 4

まず，特定のユーザが作成した UST データを 33 曲分収集する．次に，集めたデータのうち 30 曲をコーパスとし，4 章で解説した LSTM モデルで学習する．この際，エポック数は 1000，バッチサイズは 32 とする．この学習モデルを使用し，正解データを元に作成された未知データの各パラメータを予測する．なお，本実験ではベースラインとして，UTAU の既存プラグイン機能である AutoPitchwriter を使用している．

実験で評価する項目は，ポルタメントの出現箇所とポルタメントの形状の類似度の 2 点である．ポルタメントの出現箇所では，ピッチポイントが 3 個以上の抑揚をつけるポルタメントが各音に付与されているか正解データと予測データで比較し正答率，適合率，再現率，F 値を算出する．ポルタメントの形状の類似度では，正解データと予測データでポルタメントに何 cent 差があるか 0.1 ms ごとに算出し，その平均を取る．なお，ポルタメントが出現しない箇所及び出現する前後では，シフト値が 0 cent であると仮定して比較する．

ポルタメントの出現箇所の実験結果は，表 2 のとおりである．ポルタメントの形状の類似度の実験結果は，表 3 のとおりである．類似度は，値が小さいほど正解データと誤差が少なく精度が高い結果となっている．

表 2 ポルタメントの出現箇所

	Accuracy	Precision	Recall	F-measure
モデル 1	65.0	61.0	99.0	75.3
モデル 2	68.4	73.2	63.2	67.8
モデル 3	73.2	82.0	64.2	71.7
モデル 4	63.7	68.5	60.5	64.2
ベースライン	64.1	90.4	37.6	52.7

表 3 ポルタメントの形状の類似度

モデル 1	モデル 2	モデル 3	モデル 4	ベースライン
10.0	10.6	10.2	12.8	12.1

5. 考察と今後の課題

表 2，表 3 から，モデル 1，3 がモデル 2，4 に比べて出現箇所の F-measure 及び形状の類似度が高くなっていた．このことより，ポルタメントの生成を判別するパラメータをまとめて学習するモデルの方が精度が高いと言える．また，モデル 1，2 とモデル 3，4 を比較した結果，モデル 1，2 の方が精度が高くなっていた．このことより，前の 31 音を関連付ける 32 次元の LSTM モデルの方が精度が高いと言える．

本稿では，合成音声ソフトに用いられるテキストデータを元にコーパスの作成を行い，未知の楽譜のパラメータを予測するモデルを提案した．今後は，入力に歌詞やフレーズなどのデータを与えることで精度の向上を目指す．

参考文献

[1] 飴屋／菖蒲. 歌声合成ツール UTAU サポートページ. 入手先 ⟨http://utau2008.web.fc2.com/⟩ (参照 2021-12-21)

[2] HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. LSTM can solve hard long time lag problems. In: Advances in neural information processing systems. p. 473-479. 1997.