Vol.2020-MUS-126 No.14
2020/2/18

合成音声歌唱におけるヴィブラートのパラメータ自動推定

田中 瑞穂[1,a)]　竹川 佳成[1]　平田 圭二[1]

概要：近年，音声合成ソフトによる歌唱を用いた楽曲が増加している．しかし，音声合成ソフトを人間のように歌唱させるためには，声量や抑揚など，多くのパラメータを調節する必要がある．これらのパラメータ調節を，1 曲につき 1000 個近く存在する音の 1 つ 1 つに手作業で入力するのは時間がかかる．この手間がなくなれば，音声合成ソフトに歌唱させる難度が下がり，多くのユーザが使えるようになるだろう．
そこで本研究では、機械学習を用いてパラメータ調節の自動化する手法を提案する．手法としては，調節済みの音声合成ソフトのパラメータを収集する．それらのパラメータを機械学習を用いて学習する．その学習結果を元に，新たな歌唱のパラメータを自動で調節する．なお，本研究では，学習するパラメータはヴィブラートのみに限定する．

Automatic parameter vibrato of intonation in singing by vocaloid

1. はじめに

近年，VOCALOID [6] などの音声合成ソフトの普及により音声合成ソフトによる歌唱を用いた楽曲が増加している．2006 年からサービスを開始している動画投稿サイトニコニコ動画では，2020 年 1 月現在「VOCALOID」タグのついた動画が 13 年間で 53 万件投稿されている．それに伴い，音声合成ソフトもユーザに使いやすいように改良が進んでいる．2018 年 7 月には，ユーザが使いやすいよう UI が刷新された VOCALOID5 が発売されている．

しかし，このような歌唱用音声合成ソフトは，ピッチやボリュームなど様々なパラメータを調節する必要がある．パラメータは通常手動で調節するが，全ての音を手作業で調節すると時間がかかる．また，調節には決まった手順がないため，初めて音声合成ソフトを使用するユーザはどのような順序でパラメータを調節すると効率的かつ効果的か自明ではない．例として，合成音声ソフト UTAU [1] では，ヴィブラートが 7 つのパラメータで合成されているが，振幅が大きすぎると不安定に聴こえたり，周期が小さすぎるとがなり声になることがある．

これらの問題の解決策として，VocaListener [7] [8] や Sinsy [9] などの様々な歌声合成システムが開発されている．VocaListener は，ユーザの歌唱の音高や音量，声色を模倣するシステムである．歌唱による感覚的なパラメータ調節が可能であるが，歌唱が苦手なユーザは使用が困難である．Sinsy は，特定歌唱者の歌唱データを元に作成した隠れマルコフモデル (Hidden Markov Model，以下 HMM) を用いて，歌唱者の声の特徴を再現した合成音声歌唱を生成するシステムである．しかし，このモデルは 1 人の歌唱者の歌唱データを元に作成しているため，入力する楽曲がどのような曲であってもその歌唱者の特徴が出現するという問題がある．

本研究では，合成音声ソフトのヴィブラートのパラメータを機械学習で学習し，未知の楽譜にヴィブラートを自動的に付与する手法を提案する．この手法を使えば，パラメータ調節が苦手なユーザでも，自動化によりパラメータの調節が可能になる．なお，数あるパラメータの中からヴィブラートを選んだ理由は，歌唱の揺れや響きが歌声らしさを出すうえで重要な要素であり [10]，合成音声ソフトでの収集が可能なためである．

本研究では，時系列を考慮した機械学習 Long short-term memory (以下 LSTM) を使用する．時系列を考慮した機械学習は，他の音との前後関係の情報が必要な音楽の学習に適している．そのため，時系列情報に依存しているヴィブラートのパラメータの学習も可能である．また，橋本らの比較でも HMM よりも深層学習を用いた音声合成システムの方が品質が向上することが分かっている [4]．そのため，

[1] 公立はこだて未来大学
Future University Hakodate, Hakodate, Hokkaido, Japan
[a)] b1016128@fun.ac.jp

© 2020 Information Processing Society of Japan

LSTM も合成音声のパラメータの学習に適していると考えられる．

2. 関連研究

2.1 VocaListener

中野らは，合成音声ソフトの音高及び音量の調節にユーザの歌唱を用いるシステム，VocaListener [7] を開発した．これは，ユーザの歌唱を合成音声に模倣させるシステムである．それまでにも，ユーザの歌唱から音高や音量等を入力する研究は存在した．しかし，使用する合成音声システムや音源が変化すると歌唱の聴こえ方が変わってしまい，分析結果の誤りを修正する仕組みもなかった．中野らは，合成音声の歌唱と入力歌唱を近づけるために，生成した合成パラメータを一度合成音声ソフトで歌唱させてその音声から再び推定を行うことで，合成音声システムや音源の変化に対して柔軟に対応できるようにした．また，分析結果をユーザがインタラクティブに修正できる支援機能も作成した．

中野らは VocaListener の開発後，ユーザ歌唱の声色を模倣するシステム，VocaListener2 [8] を開発した．これは，VocaListener で模倣した多様な歌声を元に多次元の声色空間を構成し，ユーザ歌唱を声色を声色空間内で再現するシステムである．これにより声色の調節が可能となり，より表情豊かな合成音声歌唱が生成出来る．

これらのシステムを用いることで，ユーザは歌唱による感覚的なパラメータ調節が可能である．しかし，歌唱が苦手なユーザは使用が難しいという問題もある．また，VOCALOID 楽曲には人間の歌唱を想定していない楽曲も存在するため，そのような楽曲への使用も困難である．

2.2 Sinsy

大浦らは，特定歌唱者の歌唱データを元に作成した HMM を用いた歌声合成システムとして，Sinsy [9] を開発した．これは歌唱データの声質や声量などの特徴量をモデル化し，任意の楽譜に対応した歌声を合成するシステムである．このシステムは，平滑化されやすく正しく学習するのが困難なビブラートの専用モデルの導入，学習データに現れないコンテキストに対応するためにコンテキストクラスタリングを用いたモデルの生成などを行っている．また，少量のデータから歌唱の特徴を再現するために話者適応手法を用いている．

このシステムは，ユーザによるパラメータ調節が不要なため，パラメータ調節が苦手なユーザに対して有効である．しかし，このモデルは 1 人の歌唱者の歌唱データを元に作成しているため，入力する楽曲がどのような曲でもその歌唱者の特徴が出現するという問題がある．

2.3 歌唱のずれを生成する Hidden Markov Model (HMM)

藤田らは，音声合成ソフトのパラメータの中から，歌唱のリズムが自然になるように意図的に発声タイミングをずらすことが難しいタメ表現などの歌唱のずれに着目した．そこで，歌唱のずれを生成する HMM [2] を用いて歌唱にずれを自動的に付与するシステムを提案した．構築モデルは，音高，歌詞，拍節を隠れ状態とし，ミリ秒単位のずらし時間を出力とした．

2.4 Long short-term memory (LSTM)

LSTM [5] とは，時系列を考慮した機械学習である再帰型ニューラルネットワーク (Recurrent Neural Network，以下 RNN) の 1 種である．LSTM では，RNN で問題になる勾配爆発や消失問題を忘却ゲートを用いて対処している．そのため，長い系列のデータを学習することが可能であり，音声認識や文章生成で用いられている．

LSTM の構造を，図 1 に示す．LSTM ブロックは，忘却ゲート，入力調整ゲート，出力ゲートの 3 つのゲートを持ち，これらを組み合わせて出力と隠れ状態の計算を行う．

本研究では，図 1 の LSTM の他に，双方向 LSTM [3] を用いる．双方向 LSTM の構造を，図 2 に示す．このモデルは，時系列に沿って学習する LSTM と，時系列と逆方向に学習する LSTM の出力を統合している．これにより，各時刻の出力に全ての時刻の情報を含めることが出来る．

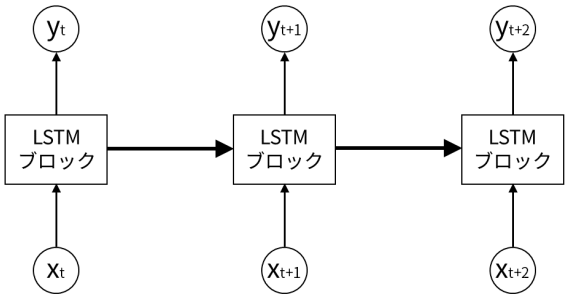

図 1 LSTM の構造

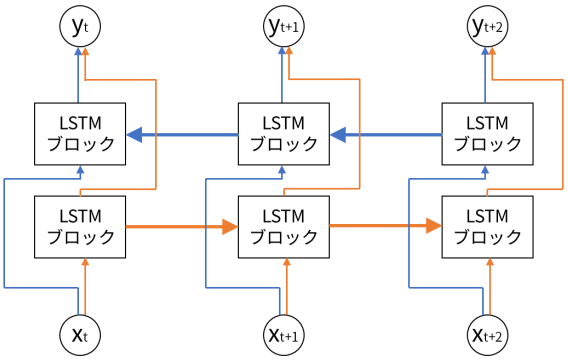

図 2 双方向 LSTM の構造

© 2020 Information Processing Society of Japan

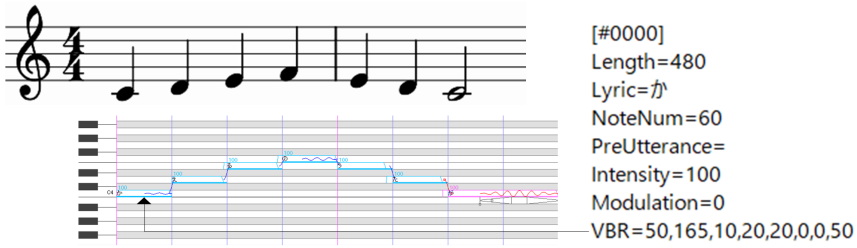

図 3 かえるのうた　楽譜 (左上)，UTAU 画面 (左下)，UST コード (右)

3. UTAU

本研究では，ヴィブラートのパラメータ収集に UTAU を用いる．

3.1 UTAU の概要

UTAU[1] とは，飴屋／菖蒲により開発されたフリーの音声合成ソフトである．フリーソフトであるため多くの楽譜 UTAU Sequence Text (以下 UST) が一般公開されており，データの収集に適している．

初めに，UST の中で学習に使用する 4 つのパラメータについて説明する．

- length：音価を示す．全音符が 1920 であり，編集画面上のマウスドラッグによる上限は 7680 (四全音符)，下限は 15 (128 分音符) である．また，手動で 0 を入力した場合，強制的に 15 に変更される．(1〜14 は入力可能)
- Notenum：音高を示す．C4 が 60 であり，編集画面上のマウスドラッグによる上限は 107 (B7)，下限は 24 (C1) である．なお，編集画面外で 24 以下の値を入力してから読み込んだ場合，エラーが発生して編集画面が開かなくなる．また，120 (C9) 以上の値を入力した場合は，異なる音高で発声する．(108〜119 は入力可能)
- Lyric：歌詞を示す．入力した歌詞に対応した音を発声するが，”R”，”r”，” ”は休符として扱われる．
- VBR：ヴィブラートコントローラを用いて作成したヴィブラートを示す．ヴィブラートコントローラについては，次節で解説する．

UTAU が UST を読み込むと，図 3 のようなピアノロールを表示する．UTAU による歌唱の調節は，この編集画面で行われる．ピアノロール上に表記される四角い図形がノートであり，横幅が音価，上下の位置が音高を示す．また，ノート上に表示されている線 (以下ピッチ線とする) は，歌唱時のピッチの変化を示している．UTAU ではこのピッチ線の形状を変更することで，ヴィブラートやポルタメントなどの抑揚を調節する．

ピッチ線の編集方法は，大きく分けて mode1 と mode2 の 2 通りがある．mode1 は，ピッチを時間ごとに指定する．mode2 は，ヴィブラートコントローラとポルタメント機能を用いる．本研究では，mode2 のヴィブラートコントローラで生成されたヴィブラートを学習に用いる．また，UTAU では自分で調節する他に，組み込みツールやプラグインによる自動調節が可能である．次項で，比較実験に用いる組み込みツール「おま☆かせ」の調節機能についても説明する．

3.2 ヴィブラートコントローラ

歌唱におけるヴィブラートとは，ある音程を中心に音高を周期的に震わせる歌唱技術である．ヴィブラートコントローラでは，ヴィブラートを図 4 のように割合，周期，深さ，入，出，位相，高さの全 7 つのパラメータで表現する．

割合は，音価に対するヴィブラート長の割合を示す．周期は，ヴィブラートの波形の周期を示す．深さは，ヴィブラートの振幅を示す．入は，ヴィブラートの深さが最大になる地点を示す．出は，ヴィブラートの深さが減少し始める地点を示す．位相は，時間軸に対するヴィブラートの波形のずれを示す．一般的な位相を 0〜360 度で示すところを，0〜100 %で示している．図 4 の位相 50 %は，180 度ずれていることを表す．高さは，見かけの音高のずれを示す．0〜100 %で入力し，1 %ごとに 1 cent 上昇する．100 %

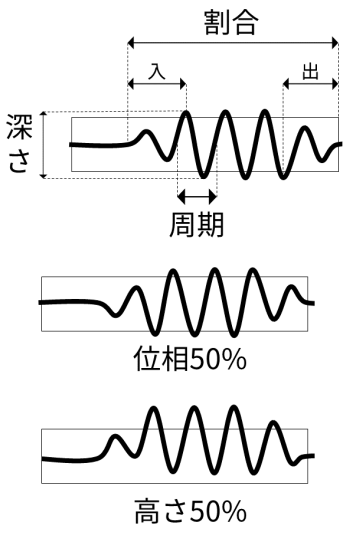

図 4 ヴィブラートの各パラメータ

© 2020 Information Processing Society of Japan

$$VBR = [PV_0, PV_1, PV_2, PV_3, PV_4, PV_5, PV_6]$$

	内容	範囲
PV_0	ヴィブラートの割合 (%)	$\{PV_0 \in Q \mid 0 \leq PV_0 \leq 100\}$
PV_1	ヴィブラートの周期 (ms)	$\{PV_1 \in Q \mid 64 \leq PV_1 \leq 512\}$
PV_2	ヴィブラートの深さ (cent)	$\{PV_2 \in Q \mid 0 \leq PV_2 \leq 200\}$
PV_3	ヴィブラートの入 (%)	$\{PV_3 \in Q \mid 0 \leq PV_3 \leq 100\}$
PV_4	ヴィブラートの出 (%)	$\{PV_4 \in Q \mid 0 \leq PV_4 \leq 100\}$
PV_5	ヴィブラートの位相 (%)	$\{PV_5 \in Q \mid 0 \leq PV_5 \leq 100\}$
PV_6	ヴィブラートの高さ (%)	$\{PV_6 \in Q \mid 0 \leq PV_6 \leq 100\}$

図 5 ヴィブラートの各パラメータの定義域

では 100 cent 上昇し，半音上の音高になる．図 4 の高さ 50 %は，50 cent 上昇していることを表す．なお，それぞれのパラメータの定義域は，図 5 のとおりである．

3.3 「おま☆かせ」調節機能

「おま☆かせ」とは，UTAU にデフォルトで搭載されている組み込みツールである．楽譜の選択部分もしくは全体に指示した調節をすることが出来る．設定できるパラメータは，ヴィブラート，母音結合，ポルタメント，モジュレーションの 4 つである．

おま☆かせでのヴィブラートの調節では，ふかさによりヴィブラートの深さ，はやさによりヴィブラートの周期，ながさによりヴィブラートの割合が設定できる．ふかさとはやさはそれぞれ 4 段階，ながさは 5 段階に分かれている．これらを組み合わせることで，全 80 種類の形状のヴィブラートを設定することが可能である．

4. 提案手法

本研究では，2 つの LSTM を用いた手法を提案する．この手法は，2 段階に分かれている．まず，ヴィブラートの出現箇所を判定する双方向 LSTM を用いて，ヴィブラートが楽譜のどの箇所に出現するか判定する．そして，ヴィブラートが出現した音は，ヴィブラートのパラメータを推定する LSTM でパラメータの値を出力する．

まず，LSTM を用いてヴィブラートのパラメータが出力出来るか確認するために，モデルを作成して予備実験を行った．作成したモデルは，2 小節分の楽譜を 32 分割して入力し，その小節の先頭に位置する音のヴィブラートのパラメータ 7 つを出力する LSTM である．しかし，このモデルを用いてヴィブラートを付与した結果，正解データに比べて極端に小さい値のヴィブラートのパラメータが出力された．これは，ヴィブラートが存在しない箇所の影響を受け，正しい学習ができなかったことが原因である．この問題を解決するために，ヴィブラートの出現箇所とパラメータの推定を別々のモデルで学習する．

ヴィブラートの付与は，図 6 の手順で行われる．まず，未知の UST から音価，音高，休符の有無を抽出する．抽出したデータは，16 分音符の大きさで分割し One-Hot ベクトルの入力データに変換する．次に，ヴィブラートの出現

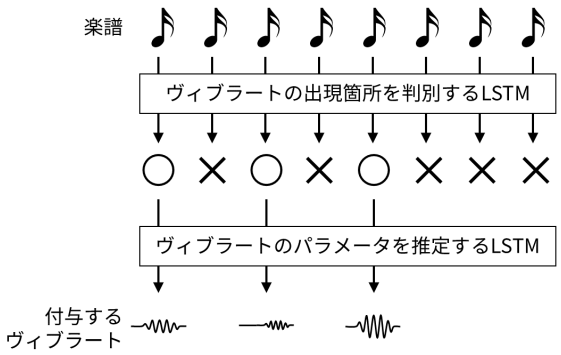

図 6 ヴィブラート付与組み合わせ LSTM モデルの概要

箇所を判別する双方向 LSTM でヴィブラートの出現箇所を 2 小節ずつ学習する．ヴィブラートと判定された音は，その音を終端として持つような 8 拍幅の旋律データに変換する．この旋律データをヴィブラートのパラメータを推定する LSTM に入力し，ヴィブラートのパラメータを連続値で出力する．最後に，予測したヴィブラートを入力に用いた UST に付与する．このように，ヴィブラートの出現箇所とパラメータの値を別々に学習することで，ヴィブラートのパラメータを推定する LSTM の学習時に不必要なヴィブラートの出現しないデータを除去することが期待される．

4.1 ヴィブラートの出現箇所を判別する双方向 LSTM

ヴィブラートの出現箇所を判別する双方向 LSTM では，2 小節分の楽譜の音高および休符を入力し，その 2 小節間のヴィブラートの出現箇所を出力する．モデルは，図 7 のとおりである．

x は入力データ，y は出力データを示している．また n は，16 分音符刻みの拍番号である ($n \geqq 0$)．入力は，図 8 のように楽譜を 16 分音符刻みにしたデータ 2 小節分である．入力層は，32 次元である．中間層は，32 次元の双方向 LSTM である．双方向 LSTM は，各時刻の出力に全て

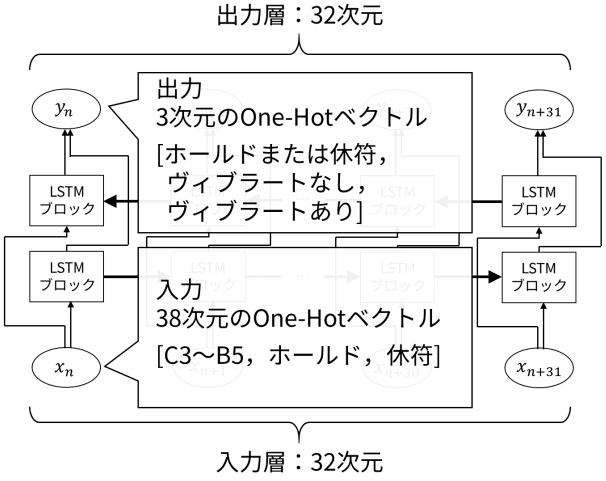

図 7 ヴィブラートの出現箇所を判別する双方向 LSTM

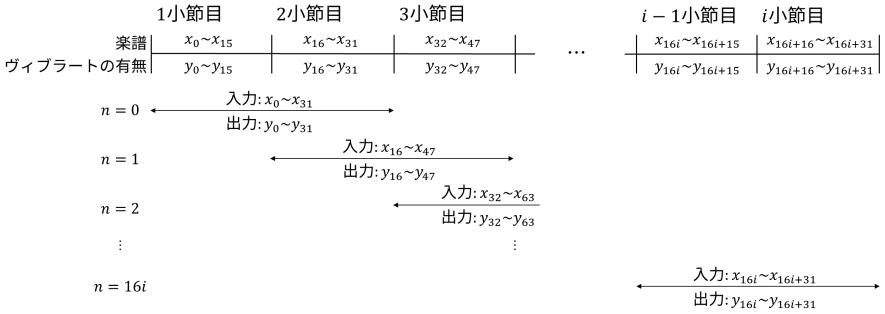

図 8 ヴィブラートの出現箇所を判別する双方向 LSTM の入出力データ

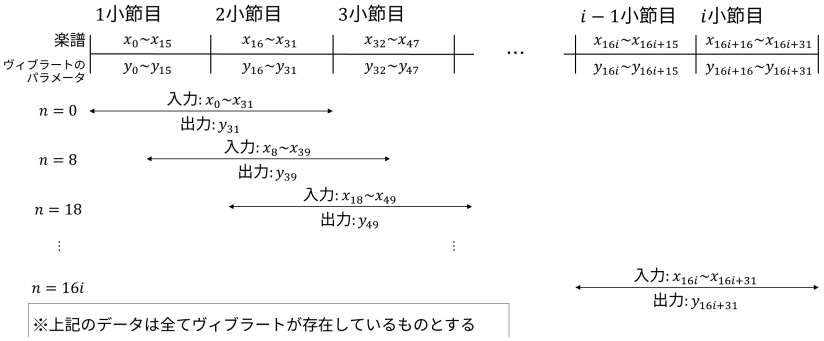

図 9 ヴィブラートのパラメータを推定する LSTM の入出力データ

の時刻の情報を含めることが出来る．出力は，図 8 のようにヴィブラートの有無を 16 分音符刻みで判定したデータ 2 小節分である．このうちヴィブラートの付与に用いるのは，1 小節目の出力のみである．出力層は，32 次元である．なお，入出力データは，1 小節ずつシフトしている．中間層の活性化関数には Tanh 関数を，出力層の活性化関数には Sigmoid 関数を用いる．また，損失関数は交差エントロピーを使用する．

4.2 ヴィブラートのパラメータを推定する LSTM

ヴィブラートのパラメータを推定する LSTM では，2 小節分の楽譜の音高および休符を入力し，1 音目のヴィブラートのパラメータ 7 つを出力する．モデルは，図 10 のとおりである．

x は入力データ，y は出力データを示している．また n は，16 分音符刻みの拍番号である．$n < 0$ の場合は，休符が入力されている．入力は，図 9 のように楽譜を 16 分音符刻みにしたデータ 2 小節分である．入力層は，32 次元である．中間層は，32 次元の LSTM である．出力は，図 9 の x_n で出現するヴィブラートのパラメータ 7 つのそれぞれの連続値である．出力層は，7 次元である．なお，入出力データは，1/16 小節ずつシフトする．ただし，ヴィブラートが付与されていないデータは，入出力データに含めずに次の箇所にシフトするものとする．中間層の活性化関数には tanh 関数を，出力層の活性化関数には恒等関数を用いる．また，損失関数は平均二乗誤差を使用する．

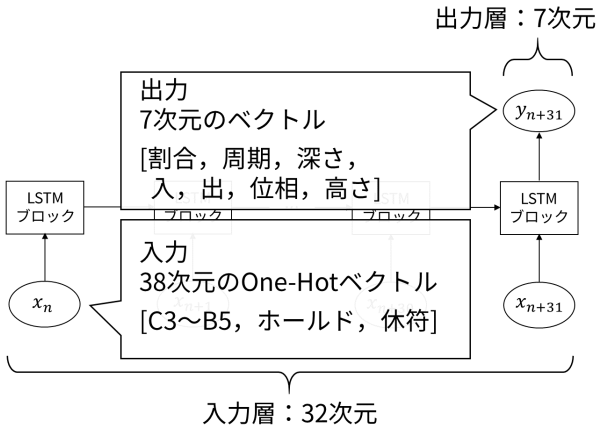

図 10 ヴィブラートのパラメータを推定する LSTM

5. 実験

本章では，4 章で提案したモデルの精度と有用性を調べるために 3 つの実験を行った．ベースラインは，おま☆かせ調節機能で調節したヴィブラートのパラメータとした．

5.1 ヴィブラートの出現箇所判別双方向 LSTM の評価実験

ヴィブラートの出現箇所判別双方向 LSTM の精度を調べる評価実験では，モデルを用いてヴィブラートを付与した楽譜と既にヴィブラートが付与された楽譜を比較し，正答率，適合率，再現率，F 値を算出した．

実験の手順は，図 11 のとおりである．正解データを

© 2020 Information Processing Society of Japan

表 1　提案手法によるヴィブラート出現箇所推定結果

楽譜	P1	P2	P3	Q1	Q2	Q3	R1	R2	R3	S1	S2	S3	T1	T2	T3
正答率	0.90	0.98	0.91	0.90	0.98	0.93	0.73	0.79	0.73	0.92	0.91	0.92	0.61	0.62	0.91
適合率	0.03	0.29	0.29	0.65	0.47	0.38	0.65	0.77	0.70	0.50	0.61	0.56	0.71	0.52	0.91
再現率	1.00	1.00	0.18	0.31	0.47	0.74	0.91	0.95	0.95	0.40	0.72	0.78	0.53	0.96	0.89
F値	0.06	0.44	0.22	0.42	0.47	0.50	0.76	0.85	0.81	0.44	0.66	0.65	0.61	0.67	0.90

表 2　従来手法によるヴィブラート付与結果

楽譜	P1	P2	P3	Q1	Q2	Q3	R1	R2	R3	S1	S2	S3	T1	T2	T3
正答率	0.62	0.57	0.53	0.32	0.38	0.41	0.72	0.72	0.86	0.64	0.64	0.62	0.62	0.74	0.82
適合率	0.70	1.00	1.00	1.00	0.98	0.99	0.63	0.67	0.86	0.88	0.85	0.79	0.77	0.74	0.84
再現率	0.30	0.15	0.14	0.08	0.06	0.15	0.95	0.96	0.97	0.97	0.47	0.43	0.86	0.90	0.85
F値	0.31	0.23	0.22	0.15	0.10	0.24	0.68	0.73	0.88	0.30	0.50	0.42	0.76	0.78	0.78

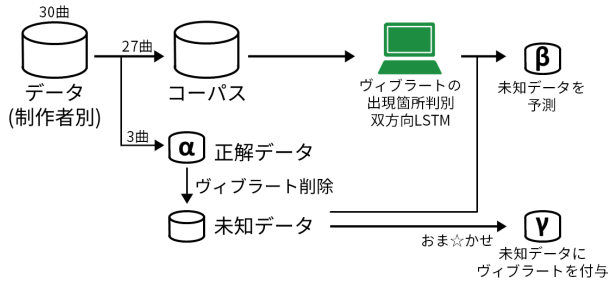

図 11　ヴィブラートの出現箇所判別双方向 LSTM の実験手順

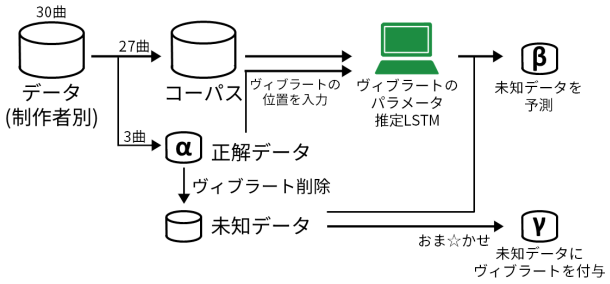

図 12　ヴィブラートのパラメータ推定 LSTM の実験手順

α，提案手法により作成したデータを β，従来手法であるおま☆かせ調節機能で作成したデータを γ とする．まず，制作者ごとに 30 曲ずつの楽譜データを収集する．本実験では，5 人分の楽譜データを収集した．集めたデータを 27 曲ずつコーパスにし，ヴィブラートの出現箇所を判別する双方向 LSTM で学習する．この際，入力データは Krumhansl-Schmuckler [11] を用いて調を揃えた．また，学習時のエポック数は 3000，バッチサイズは 2048 とした．この学習モデルを使用して，未知データのヴィブラート出現箇所を予測し，データ β とする．β とは別に，ベースラインとしておま☆かせ調節機能を用いてヴィブラートを付与したデータ γ を用意する．このように作成したデータ β とデータ γ を，正解データ α とヴィブラートの出現箇所を比較し，正答率 (Accuracy)，適合率 (Precision rate)，再現率 (Recall rate)，F 値 (F-measure) を算出する．

ヴィブラートの出現箇所判別双方向 LSTM の評価実験の結果は，表 1 および表 2 のとおりである．UST の制作者を P，Q，R，S，T として，曲にはそれぞれ 1 から 3 の番号をふっている．また，提案手法と従来手法の正答率などを比較して，値が大きい方を赤で塗っている．提案手法を用いた結果，P，Q，S は全ての楽譜において 90 %以上の正解率が出ており，従来手法の正答率を大幅に上回っていた．しかし，R，T は正答率が 60〜80 %のものがあり，楽曲によっては従来手法の正答率を下回っていた．また，提案手法は全体的に再現率の値が適合率より大きく，反対に従来手法は適合率の値が再現率より大きかった．

5.2　ヴィブラートのパラメータ推定 LSTM の評価実験

ヴィブラートのパラメータ推定 LSTM の精度を調べる評価実験では，モデルを用いてヴィブラートを付与した楽譜を既にヴィブラートが付与された楽譜と比較してヴィブラートの平均二乗誤差を求める．

実験の手順は，図 12 のとおりである．正解データを α，提案手法により作成したデータを β，従来手法であるおま☆かせ調節機能で作成したデータを γ とする．まず，5.3 節と同様に制作者 5 人の楽譜データを 30 曲ずつ収集した．次に，集めたデータを 27 曲ずつコーパスにし，ヴィブラートのパラメータ推定 LSTM で学習する．なお，ヴィブラートが含まれていないデータは，コーパスから除去する．入力データは，Krumhansl-Schmuckler [11] を用いて調を揃えた．また，出力データは，Z 得点を用いて標準化した．学習時のエポック数は 3000，バッチサイズは 2048 とした．この学習モデルを使用し，未知データのヴィブラートのパラメータを予測し，データ β に付与した．なお，この際ヴィブラートの出現箇所は，正解データと同様の箇所に設定する．β とは別に，おま☆かせ調節機能を用いてヴィブラートのパラメータを付与したデータ γ を用意する．このように作成したデータ β とデータ γ を，正解データ α とヴィブラートのパラメータを比較し，平均二乗誤差を算出する．平均二乗誤差を使用した理由は，損失関数などでも用いら

© 2020 Information Processing Society of Japan

表 3　ヴィブラートのパラメータ推定 LSTM の実験結果

楽譜	P1	P2	P3	Q1	Q2	Q3	R1	R2	R3	S1	S2	S3	T1	T2	T3
提案手法	0.004	1E-04	7E-05	0.256	0.319	0.124	0.004	0.002	0.01	0.095	0.124	0.065	0.15	0.272	0.15
従来手法	0.628	0.595	0.640	0.594	0.705	0.589	1.038	0.908	0.614	0.727	0.720	0.679	0.938	0.714	0.864

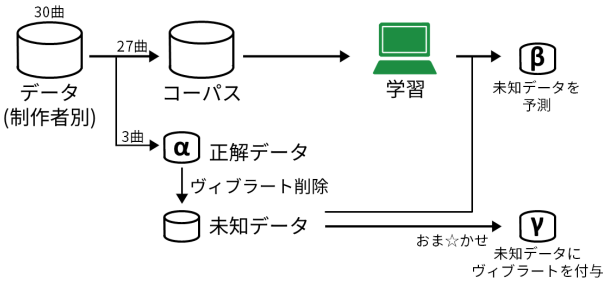

図 13　聴取実験の音源生成

れており，2 つのデータの差を求めるのに適しているためである．また，出力データを正規化したことにより範囲も統一されているためである．

ヴィブラートのパラメータ推定 LSTM の評価実験の結果は，表 3 のとおりである．全体的に従来手法の誤差よりも提案手法の誤差が下回っていた．特に P と R の提案手法の誤差が小さく，従来手法と比較して大きな差が見られた．対して Q と T の提案手法の誤差は大きく，従来手法と比較して大きな差が見られなかった．

5.3　聴取実験

聴取実験では，提案手法と従来手法で生成したヴィブラートを比較して，どちらが歌唱に良い影響を与えているか被験者に聴き比べてもらった．まず，図 13 のように提案手法と従来手法を用いてヴィブラートを付与した音声を用意する．被験者は，この音声を 5 曲分聴いて，歌唱の上手さ，感情の込められ方，歌唱の違和感の無さについて 5 点満点で評価する．なお，5 曲は全て異なる制作者の楽曲であり，モデルの学習も制作者ごとに行っている．

今回の実験では，計 18 名の被験者に音声を聴いてもらった．なお，UTAU の原音には既に感情がこもっているものが存在し，それが実験のノイズになる可能性がある．そのため，原音ファイルはノイズが少ないであろう UTAU デフォルト音声 Ver1.2 を使用した．また，音声を聴く順番が結果に影響しないようにするため，提案手法と従来手法を聴く順番をランダムに入れ替えた．

聴取実験の結果は，図 14 のとおりである．P 以外の楽曲で歌唱の上手さ，感情の込められ方，歌唱の違和感の無さの 3 項目とも提案手法よりも良いという結果が得られた．また，提案手法と従来手法を t 検定した結果，Q と R の歌唱の上手さと R の歌唱の違和感の無さに有意差が認められた ($p < 0.05$)．

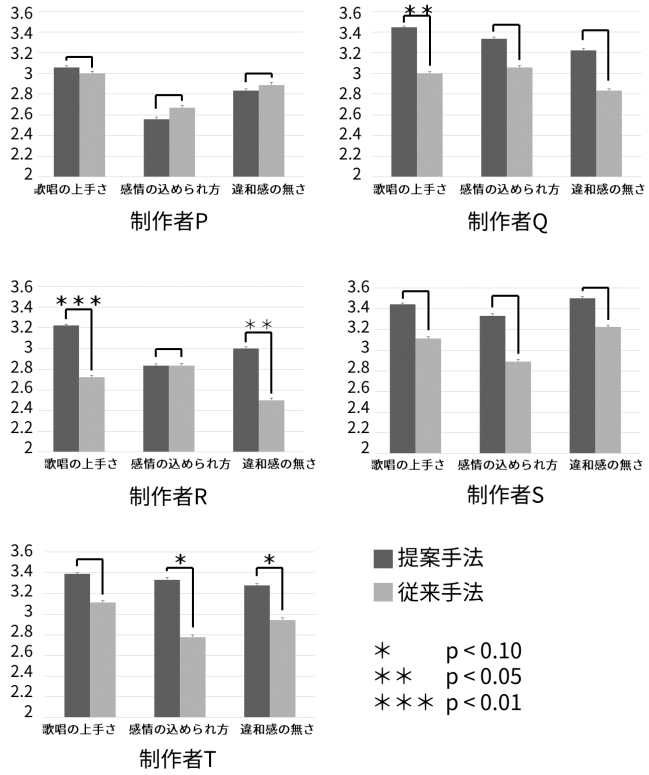

図 14　聴取実験の実験結果

6.　考察

ヴィブラートの出現箇所を予測する双方向 LSTM は，制作者 R と T の正答率が低かった (表 1，2)．これは，制作者 R と T が拗音，促音，ブレス等にはヴィブラートを付与しない，テンポの変化によってヴィブラートを削除する等，音価や音高以外の要素からヴィブラートを付与するか判断していたためではないかと考えられる．また，従来手法は適合率の値が大きく，提案手法は再現率の値が大きかった．これは，従来手法が全体にまんべんなく多くのヴィブラートを付与するのに対し，提案手法は重要な箇所にのみヴィブラートを付与していたからだと考えられる．

ヴィブラートのパラメータを予測する LSTM は，従来手法と比較して大幅に改善されていた (表 3)．しかし，制作者 Q や T のような様々なヴィブラートを使用している楽譜に対してはまだ精度が低かった．これは，学習データの数を増やし様々なパターンを学習させることで対応出来るようになると考えられる．

聴取実験では，歌唱の上手さ，感情の込め方，違和感の無さの 3 項目において P 以外の楽曲は提案手法よりも良い

© 2020 Information Processing Society of Japan

という結果が得られた (図 14)．しかし，有意差な差は一部でしか見られなかった．これはヴィブラートのみを付与するシステムでは，上記の 3 つの項目に与える影響が小さいからではないかと考えらえる．また，学習させたデータのヴィブラートの箇所がそもそも適切ではなかった可能性もあげられる．

7. おわりに

本研究では，機械学習を用いて音声合成歌唱のヴィブラートを付与するモデルの作成を行った．モデルは，ヴィブラートの出現箇所を判別する双方向 LSTM とヴィブラートのパラメータを推定する LSTM の 2 つを作成し，それを組み合わせる形で未知の楽譜のヴィブラートを予測した．この LSTM を 2 つ用いたモデルの精度を検証するために，評価実験と聴取実験を行った．評価実験では，機械学習を用いて予測する提案手法と合成音声ソフトに組み込みツールによる従来手法の 2 つの比較をそれぞれの LSTM で行った．結果として，促音やブレスのような特殊な音が考慮されていないこと，ヴィブラートのパターンが多い場合の精度が低いことなどが判明した．聴取実験では，被験者に提案手法と従来手法で作成した音源を聴いてもらい，歌唱の上手さや感情表現が出来ているか，違和感の無さなどを比較した．結果として，提案手法は従来手法と比較して良い結果が出たが，有意な差はほとんど見られなかった．

今後の課題は，大きく分けて 3 つある．1 つ目は，ヴィブラートを学習するモデルの精度を上げることである．具体的な方法としては，特殊な音やテンポを入力に加えること，調を揃える方法を変更することなどが挙げられる．また，モデルの層数や次元数も検討する必要がある．2 つ目は，ヴィブラート以外のパラメータを学習することである．現在は，ポルタメントの学習を検討している．3 つ目は，人間の歌唱によるヴィブラートと UST から収集したヴィブラートの 2 つをそれぞれ学習して，どちらが合成音声に対して適切なヴィブラートなのか比較することである．従来手法は，人間の歌唱を元にヴィブラートを学習するシステムが多い．しかし，UST から収集したヴィブラートに十分な表現力があれば，人間の歌唱データを扱うよりも効率的にヴィブラートを推定できるのではないかと考えられる．

謝辞

研究を通じて議論をしていただいた寺井あすか先生（公立はこだて未来大学）に感謝いたします．本研究は JSPS 科研費 16H01744 の助成を受けたものです．

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/⟩, (2020.01.27).
[2] 藤田千尋，竹川佳成，平田圭二：歌唱における発声タイミングのずらし時間抽出と発声タイミングモデルの提案，第 80 回全国大会講演論文集，pp.137-138（2018）．
[3] Graves, A and Jürgen, S.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural networks, Vol.18, No.5-6, pp.602-610 (2005).
[4] 橋本佳，高木信二，深層学習に基づく統計的音声合成，日本音響学会誌，Vol.73，No.1，pp.55-62（2017）．
[5] Hochreiter, T and Jürgen, S.: LSTM can solve hard long time lag problems, In Advances in neural information processing systems, pp.473-479 (1997).
[6] 剣持秀紀，大下隼人：歌声合成システム VOCALOID，情報処理学会研究報告 2007-MUS-72，pp.25-28（2007）．
[7] 中野倫靖，後藤真孝：VocaListener: ユーザ歌唱の音高および音量を真似る歌声合成システム，情報処理学会論文誌，Vol.52，No.12，pp.3853-3867（2011）．
[8] 中野倫靖，後藤真孝：VocaListener2: ユーザ歌唱の音高・音量に加えて声色変化も真似る歌声合成システム，情報処理学会論文誌，Vol.54，No.6，pp.1771-1783（2013）．
[9] 大浦圭一郎，間瀬絢美，山田知彦，徳田恵一，後藤真孝：Sinsy:「あの人に歌ってほしい」 をかなえる HMM 歌声合成システム，研究報告音楽情報科学，pp.1-8（2010）．
[10] 齋藤毅，辻直也，鵜木祐史，赤木正人：歌声らしさの知覚モデルに基づいた歌声特有の音響特徴量の分析，日本音響学会誌，Vol.64，No.5，pp.267-277（2008）．
[11] Temperley, D.: What's key for key? The Krumhansl-Schmuckler key-finding algorithm reconsidered, Music Perception: An Interdisciplinary Journal, vol.17, no.1, pp.65-100 (1999).

© 2020 Information Processing Society of Japan