Vol.2021-MUS-130 No.25
Vol.2021-EC-59 No.25
2021/3/17

合成音声歌唱のポルタメントの統計的性質に基づくLSTMに入力する特徴量の検討

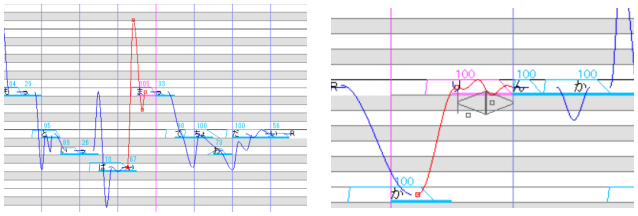
田中 瑞穂[1,a)]　竹川 佳成[1]　平田 圭二[1]

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．それに伴い，人間歌唱を模倣する歌声自動合成システムが開発されている．しかし，合成音声歌唱には，歌唱の平坦さを軽減するためにユーザが生み出した固有の歌唱技術（急なピッチ変化や短いヴィブラートなど）が存在する．これらの歌唱技術は人間が歌唱することが想定されていないため，従来の人間歌唱を対象とした歌唱モデルでは効率よく学習，推定することが難しいと考えられる．そこで本研究では，UTAU のポルタメントについて統計を取り，その統計結果から特徴量を検討し，音声合成歌唱の表情付けのためのパラメータを LSTM を使用して学習，推定する．ポルタメントの統計では，各要素ごとのポルタメントの出現率とポルタメントの要素の傾向について調べた．その結果，音のタイプ，音高，音価がポルタメントの出現率に影響を与えていることがわかった．また，ポルタメントに用いる線の形状は重要視されていなかった．これらの統計的性質の観察結果から，学習に用いる特徴量を決定し，LSTM モデルで学習，推定を行う．このモデルを使用してポルタメントを付与した音声と従来手法で付与した音声の比較実験を行ったところ，抑揚の有無と歌唱の自然さ共に従来手法の方が高い評価を得た．しかし，一部楽曲の歌唱の自然さは，提案手法の方が高い評価を得た．

1. はじめに

近年，VOCALOID [1] などの音声合成ソフトの普及により合成音声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，2021 年 2 月現在 VOCALOID のタグがついた動画が 14 年間で 57 万件投稿されている．このような音声合成ソフトの普及に伴い，合成音声歌唱に人間歌唱を模倣させる動きが広まっている．例として，VocaListener [2][3] や Sinsy [4] などの歌声合成システム，NEUTRINO [5] や Cevio AI [6] などの音声合成ソフトが挙げられる．これらは全て，人間歌唱を模倣することで合成音声歌唱を人間歌唱に近づけている．

しかし，合成音声歌唱には，図 1 のような固有の歌唱技術が存在する．これらの歌唱技術は，合成音声特有の平坦な歌唱に抑揚をつけるために生み出されたものである．音声合成ソフトを使用するユーザは，これらの歌唱技術を各種パラメータを調節して表現することで，音素の少ない音声ライブラリでも幅広い歌唱表現をさせることが可能である．しかし，これらの特殊な歌唱技術は人間が歌唱することが想定されていないため，従来の人間歌唱を模倣する手法では再現が困難である．そこで，本研究ではユーザが調節した合成音声歌唱のパラメータを模倣することで，合成音声歌唱固有の歌唱技術を再現する．



(a) 急なピッチ変化　(b) 短い音へのヴィブラート付与

図 1 合成音声歌唱固有の歌唱技術

本研究では，音声合成ソフト UTAU[7] のポルタメントのパラメータを機械学習で学習し，未知の楽譜のポルタメントのパラメータを予測する手法を提案する．一般的にポルタメントとは，音から音に移る際に滑らかに音程を変える演奏及び歌唱技法のことを示す．対して UTAU のポルタメントは，音と音を滑らかに繋げる他，歌唱のピッチを途中で変化させて歌唱表現の幅を広げる役割を担う．その役割から UTAU のパラメータの中でも重要度が高く，合成音声歌唱固有の歌唱技術の表現にも使用される．よって，本研究ではポルタメントに焦点を当てて学習を行う．なお，ポルタメントの学習に用いる特徴量を決定するため

[1] 公立はこだて未来大学
Future University Hakodate, Hakodate, Hokkaido, Japan
[a)] g2120026@fun.ac.jp

に，4 章でポルタメントについての統計を行う．

本研究では，時系列を考慮した機械学習 Long short-term memory [8] (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 [9]．そのため，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．

本論文の構成は，以下のとおりである．2 章では，関連研究と比較し，本研究の位置付けを明確にする．3 章では，研究に使用する音声合成ソフト UTAU と UTAU で使用されるポルタメントについて解説する．4 章では，ポルタメントの学習に用いる特徴量を決定するために，ポルタメントに関連する要素の統計を行う．5 章は，LSTM モデルと 4 章の結果から決定した特徴量について説明する．6 章では，5 章で提案したモデルの聴取実験を行う．7 章で実験結果の考察を行い，8 章で今後の展望をまとめる．

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 [2] 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 [3] 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，このシステムは歌唱が苦手なユーザの使用が困難である．また，合成音声歌唱固有の歌唱技術は人間歌唱が困難なものが多いため，人間歌唱を模倣するシステムでは表現が難しいという問題がある．

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 [4] 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 本研究の位置付け

本研究では，固有の歌唱技術を合成音声歌唱のパラメータから学習し，歌唱表現の幅を広げるシステムを作成する．合成音声歌唱の歌唱表現に幅を持たせるための手法として，人間歌唱の模倣や声色の異なる音声ライブラリの作成が挙げられる．しかし，これらの手法は別途音声データが必要であるため，音声データが追加できない音源での使用が困難である．対して，ユーザが音声ライブラリを生成する音声合成ソフトでは，音素の少ない音声ライブラリでも歌唱の表現に幅を持たせるため，固有の歌唱技術が生み出されている．本研究では，固有の歌唱技術を記号レベルの合成音声歌唱のパラメータから学習，推測することで，少量のデータから自動で歌唱表現の幅を広げるシステムの作成を目指す．

3. UTAU

本研究では，音声合成ソフト UTAU で使用されるパラメータをコーパスに用いる．

3.1 UTAU の概要

UTAU[7] 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) をユーザが一般公開しており，データの収集に適している．

はじめに，UST の中で学習に使用するパラメータについて解説する．

- 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，数値入力で 0 を指定した場合，強制的に 15 に変更される． (1〜14 は入力可能)

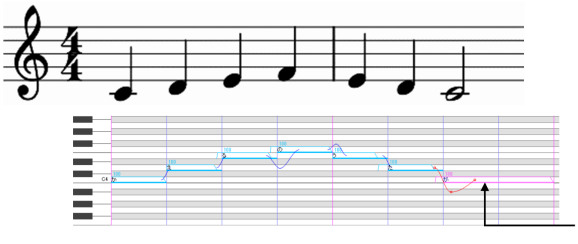


図 2 かえるのうた　楽譜 (左上)，UTAU 画面 (左下)，一部省略した UST コード (右)

- 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．
- PBS, PBW, PBY, PBM : ポルタメントの構成要素である．ポルタメントについては，3.2 節で解説する．

UTAU が UST を読み込むと，図 2 のようなピアノロールを表示する．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，合成音声歌唱の抑揚を表現する．

ピッチ線の編集方法は，大きく分けて mode1 と mode2 の 2 種類がある．mode1 は，一定時間ごとにピッチを指定するピッチベンドを調節してピッチ線の形状を変更する．mode2 は，ポルタメントで生成したピッチ線をピッチポイントと呼ばれる可動点を増やすことで形状を変更する．本研究では，mode2 から生成したピッチ線を学習に用いる．なお，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を使用して一部パラメータ調節を自動化することができる．本研究では，カノンにより開発された UTAU 用プラグイン AutoPitchwriter を比較実験に用いる．

3.2 ポルタメント

音楽用語におけるポルタメントは，音から音に移る際に滑らかに音程を変える演奏及び歌唱技法を示す．対して UTAU のポルタメントは，音を滑らかに繋げるだけではなく，歌唱に抑揚をつける機能として用いられる．

UTAU でポルタメントと呼ばれる機能は，図 3 の正方形で示されているピッチポイント (ピッチ可動点とも呼ばれる) の間に指定した形状のピッチ線を生成する．デフォルトのポルタメントはピッチポイントが 2 点のみで，効果も音を滑らかに繋げるのみである．このピッチポイントを増やしてピッチ線を曲げることで，音程の変化が必要なしゃくりやオーバーシュート，プレパレーションを表現することができる．なお，本稿では，2 点のピッチポイントで構成された一般的なポルタメントを通常のポルタメント，3 点以上のピッチポイントで構成された抑揚をつけるポルタメントを複雑なポルタメントと定義する．

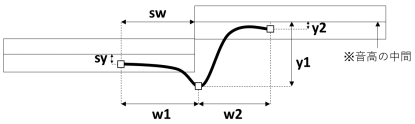


図 3 ポルタメントの各パラメータ

表 1 ポルタメントの各パラメータの定義域

$$\begin{aligned}\mathrm{PBS} &= [sw, sy]\\ \mathrm{PBW} &= [w_1, w_2, w_3 \ldots]\\ \mathrm{PBY} &= [y_1, y_2, y_3 \ldots]\\ \mathrm{PBM} &= [m_1, m_2, m_3 \ldots]\end{aligned}$$

	内容	範囲
sw	ポルタメントの始点 (ms)	$\{sw \in Q \mid -\text{前音の音価} \leq sw \leq 0\}$
sy	ポルタメントの始点 (cent)	$\{sy \in Q\}$ ※前音が休符以外の時は$sy=0$
w_n	ピッチポイントの間隔 (ms)	$\{w_n \in Q \mid 0 \leq w_n \leq \text{前音+現音の音価}\}$
y_n	ピッチポイントのシフト値 (cent)	$\{y_n \in Q \mid -200 \leq yn \leq 200\}$
m_n	ピッチポイント間の線の形状	$\{m_n \in \{" ", s, r, j\}\}$ ※" " = 曲線 , s = 直線型 , r = R型 , j = J型

USTのポルタメントは，以下の 4 つのパラメータから構成されている．

- PBS : ポルタメントの始点の座標を示す．PBS は，現在の音の始めとポルタメントの始点の差を示す sw と前の音との音高の中心との差を示す sy から構成されている．sw の単位は ms であり，編集画面上のマウスドラッグによる上限は 0 ms，下限は前の音の音価までである．sy の単位は cent であり，編集画面上のマウスドラッグによる上限は 200 cent，下限は-200 cent である．なお，前の音が休符以外の場合は，sy は 0 となる．
- PBW : 各ピッチポイントの間隔を示す．単位は ms であり，編集画面上のマウスドラッグによる上限は前の音と現在の音の音価の総計まで，下限は 0 ms である．
- PBY : 各ピッチポイントのシフト値を示す．単位は cent であり，編集画面上のマウスドラッグによる上限は 200 cent，下限は-200 cent である．なお，最後のピッチポイントのシフト値は，休符以外の場合は 0 と

なる．

- PBM：ピッチポイント間の線の形状を示す．形状は，曲線，直線，R 型 (r に似た形状の線)，J 型 (j に似た形状の線) の全 4 種類である．指定がない場合は，全て曲線型となる．

4. ポルタメントの統計的性質

学習前に，ポルタメントの傾向を調べるための統計を取る．統計内容は，大きく分けて 2 種類である．1 つ目は，ポルタメントの出現率である．これは，音のタイプや音の高さ，音の長さがポルタメントの出現率に与える影響を分析したものである．2 つ目は，ポルタメントの構成要素の分布である．これは，3.2 節で説明したポルタメントの 4 つの要素の傾向を分析したものである．なお，統計に使用したデータは，61 名のユーザによって作成された 1 つ以上のポルタメントを含む UST1584 曲である．

4.1 ポルタメントの出現率

本項では，音のタイプ，音価，音高でサンプルを分けて通常のポルタメントを含むポルタメントの出現率と複雑なポルタメントの出現率について統計を取る．

まず，音のタイプごとのポルタメントの出現率を示す (図 4)．音のタイプは，通常の音，休符，ブレスのような特殊な音の 3 つに分けている．休符は，ポルタメントの出現率が 1.3 %，特殊なポルタメントの出現率が 0.1 %とほとんどの場合で出現していない．また，通常の音の場合 89.5 %，特殊な音の場合 65.9 %にポルタメントが出現している．ただし，前の音が休符の場合や前の音と音高の変化がない場合は，ポルタメントが出現しないことが多い．なお，複雑なポルタメントは通常の音の場合 25.3 %，特殊な音の場合 31.8 %と特殊な音の方がやや出現率が高くなっている．

次に，音高ごとのポルタメントの出現率を示す (図 5)．音高の範囲はノートナンバー 36 (C2) から 95 (B6) であり，1 オクターブごとに分けて集計している．なお，休符のサンプルは除いて統計を取っている．ポルタメントの出現率は，ほぼすべての音高で出現率が 90 %前後である．一方複雑なポルタメントの出現率は，60(C4) から 83(B5) の頻繁に使用する音高で低くなっている．

次に，音価ごとのポルタメントの出現率を示す (図 6)．音価は，16 分音符以下，8 分音符以下，4 分音符以下，2 分音符以下，全音符以下，全音符より大きい音価に分けている．音高と同様に休符のサンプルは除いている．ポルタメントと複雑なポルタメント共に音価が大きいほど出現率が高い．特に複雑なポルタメントの出現率は，16 分音符以下の場合 20%，全音符以上の場合 38%とポルタメントの出現率よりも大きな差がある．

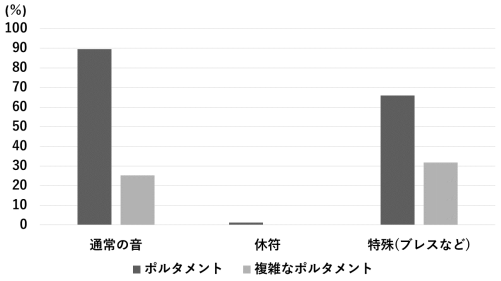


図 4 音のタイプごとのポルタメントの出現率

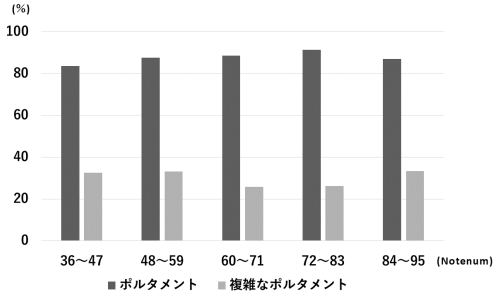


図 5 音高ごとのポルタメントの出現率

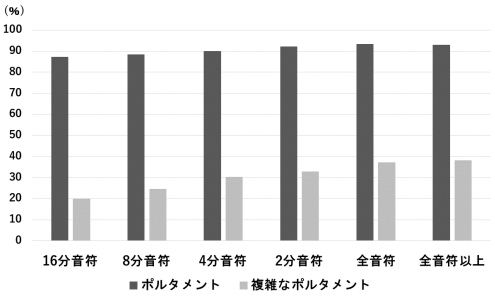


図 6 音価ごとのポルタメントの出現率

4.2 ポルタメントの構成要素の分布

本項では，各構成要素の傾向について統計を取る．統計を取る構成要素は，ポルタメントの開始地点，ポルタメントの終了地点，ピッチポイントのシフト値，ピッチポイント間の線の形状の 4 つである．

まず，ポルタメントの開始地点を示す (図 7)．これは，前の音の音価に対して何 %の地点から音高の変化が始まるか統計を取ったものである．最頻値は 85 %であり，70 %から 90 %が開始地点のサンプルが全体の半分を占めている．

次に，ポルタメントの終了地点を示す (図 8)．これは，現在の音の音価に対して何 %の地点で音高の変化が終わるか統計を取ったものである．最頻値は 21 %であり，10 %から 30 %が開始地点のサンプルが全体の半分を占めている．また，終了地点が 99 %を上回るサンプルが全体の 5 %を占めている．

次に，ピッチポイントのシフト値を示す (図 9)．これは，複雑なポルタメントで抑揚をつける際に何 cent シフトするか統計を取ったものである．グラフは縦軸の個数が対数である．最頻値は 0cent であり，全体の 55 %である．0

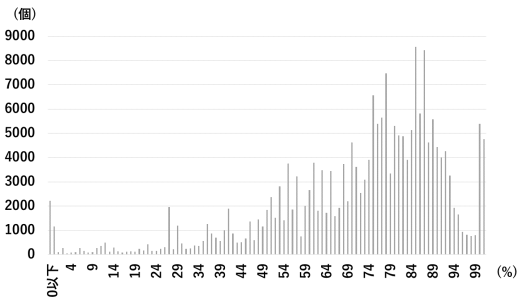


図 7　ポルタメントの開始地点の分布

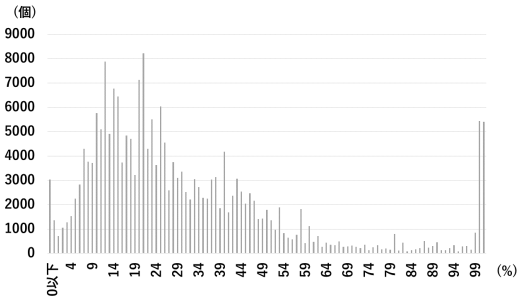


図 8　ポルタメントの終了地点の分布

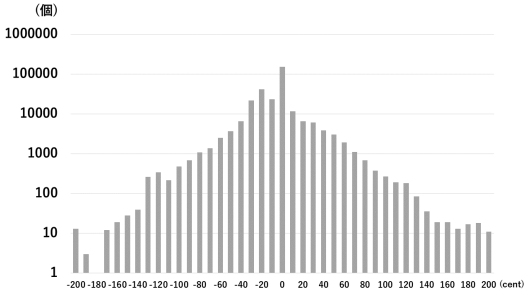


図 9　ピッチポイントのシフト値の分布

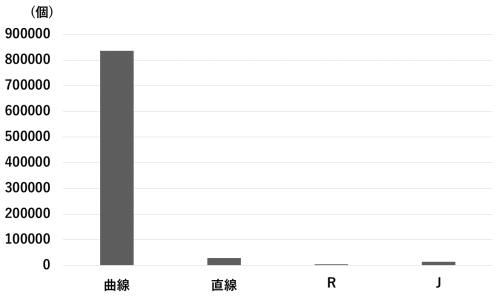


図 10　ピッチポイントのシフト値の分布

cent を除いた場合の最頻値は-20 cent であり，次いで-10 cent，-30 cent が多い．0 cent を除いてシフト値をマイナスとプラスに分けた場合，シフト値がマイナスで音高が下がっているサンプルが全体の 74.1 %を占めている．また，特殊な歌唱技法としてを 1 オクターブ以上の急なピッチ変化を紹介したが，半オクターブ以上シフトするサンプルは 2.6 %，1 オクターブ以上シフトするサンプルは全体の 0.2 %である．

次に，ピッチポイント間の線の形状を示す (図 10)．これは，PBM で指定する全 4 種の線の形状について統計を取ったものである．曲線が全体の 94.6 %を占めている．

4.3　統計的性質の考察

出現率について考察する．音のタイプごとの統計結果では，休符は前の音との繋がりを作る必要がないためポルタメントが出現しない．ただし，前の音の終わりを処理するために，休符にポルタメントを付与しているサンプルが 1 %程度存在している．また，ブレスなどの特殊な音は，通常の音よりも複雑なポルタメントの出現率が高い．これは，特殊な音はフレーズの先頭になることが少なく，複雑なポルタメントが付与しやすいためだと考えられる．音高ごとの統計結果では，音高が C4 から B5 のユーザが頻繁に使う音は，複雑なポルタメントの出現率が低い．これは，ユーザが頻繁に使う音は抑揚の癖が少ないためだと考えられる．音価ごとの音のタイプごとの統計結果では，音価が大きい音は，ポルタメントと複雑なポルタメント共に出現率が高い．これは，音が長くなるほどピッチポイントを増やす余地ができるためだと考えられる．

構成要素についての考察は，以下のとおりである．ポルタメントの開始地点の統計結果では，70 %から 90 %が全体の半分を占めている．また，ポルタメントの終了地点の統計では，10 %から 30 %が全体の半分を占めている．これは，ポルタメントのデフォルト値が開始地点 80 %及び終了地点 20 %であるため，デフォルト値の付近に分布しているといえる．また，ポルタメントの終了地点が 99 %を上回るサンプル全体の 5%を占めている．これは，音全体に抑揚をつけて終了地点が音の最終地点になったためだと考えられる．ピッチポイントのシフト値の統計結果では，0 cent を除いてからマイナスとプラスに分けた場合，シフト値がマイナスで音高が下がっているサンプルが全体の 74.1 %を占めている．このことより，音高を下げるしゃくりやプレパレーションが音高を上げるオーバーシュートよりも重要視されていると考えられる．ピッチポイント間の線の形状の統計結果では，曲線が全体の 94.6 %を占めている．このことより，線の形状を重要視しているユーザは少ないと考えられる．

この統計的性質の観察結果から，学習に用いる特徴量の選択を行う．音のタイプ，音価，音高はすべてポルタメントの出現率に影響を与えているため，この 3 つの要素は入力データに用いる．また，ポルタメントの構成要素は，ピッチポイント間の線の形状は重要視されていなかったため，ピッチポイントの座標のみを出力データに用いる．

5.　提案手法

5.1　学習モデル

学習モデルでは，32 音分の音価，音高，音のタイプを入力し，各音のピッチポイントの座標を出力する．モデルは，

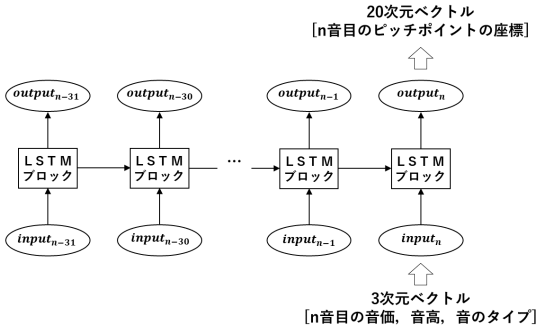


図 11 ピッチポイントを予測する LSTM

表 2 入力データの構成
$input = [i_0, i_1, i_2]$

	内容	範囲
i_0	inputの音価 (1920tick=1)	$\{i_0 \in Q \mid 0 < i_0\}$
i_1	inputの音高	$\{i_1 \in Q \mid 0 \leq i_1 \leq 1\}$
i_2	音のタイプ	$\{i_2 \in \{0,1,2\}\}$ ※0 = 休符 1 = 特殊な音(ブレスなど) 2 = 通常の音

図 11 のとおりである．

input は入力データ，*output* は出力データを示している．入力は，音価，音高，音のタイプから構成される 3 次元のベクトルである．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，10 点のピッチポイントの座標から構成される 20 次元のベクトルである．出力層は，32 次元である．なお，入出力データは，1 音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．

5.2 入力データ

入力データは，各音の音価，音高，音のタイプの 3 つの特徴量から構成されている (表 2)．音価の単位は tick であり，1920 tick を 1 として扱う．音高は UST のノートナンバーを元にしており，ノートナンバー 0 (C0) からノートナンバー 120 (C9) を 0 から 1 として扱う．音のタイプは，休符，ブレスなどの特殊な音，通常の音の 3 つに分かれており，それぞれ 0，1，2 として扱う．

5.3 出力データ

出力データは，10 点のピッチポイントの座標から構成されている (表 3)．1 点のピッチポイントを表現するために，音高のシフト値を示すベクトルと時間を示すベクトルが必要になる．音高のシフト値を示すベクトルは，1000 cent を 1 として扱う．1 つ目のピッチポイントは前の音との音高差，それ以外のピッチポイントは現在の音との音高差を用いる．時間を示すベクトルは，絶対的な表現と相対的な表現の 2 種類を用意している．これは，ピッチポイントの出現箇所が音価によって値が変わるべきかどうか確認するためである．絶対的な表現では，1000 tick を 1 として扱い，現在の音の先頭を 0 とした差を用いる．相対的な表現では，入力は全て 0 から 1 となる．1 つ目のピッチポイントは開始地点，最後のピッチポイントは終了地点を示す．それ以外のピッチポイントはポルタメントの開始地点を 0，終了地点を 1 とした場合どの地点にあるかを示す．

表 3 出力データの構成
$output = [o_{x0}, o_{y0}, o_{x1}, o_{y1}, o_{x2}, o_{y2} \ldots o_{x9}, o_{y9}]$

	内容	範囲
o_x	時間の座標(1000tick=1)	・絶対的な表現 $\{o_x \in Q \mid$ 前音の音価 $\leq o_x \leq$ 現音の音価$\}$ ・相対的な表現 $\{o_x \in Q \mid 0 \leq o_x \leq 1\}$
o_y	シフト値の座標(1000cent=1)	$\{o_y \in Q\}$

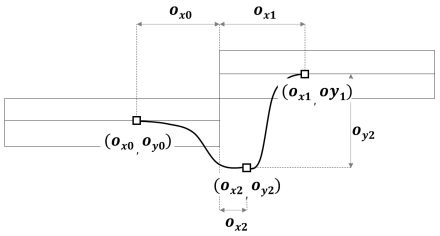


図 12 絶対的な表現

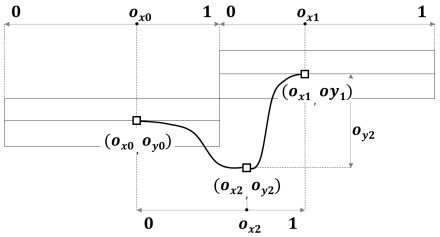


図 13 相対的な表現

出力データでは，これらのベクトルを 1 点目のピッチポイント，最後のピッチポイント，それ以外のピッチポイントという順番に並べ替えて使用している．なお，ピッチポイントが 10 点に満たない場合は，0 で埋めてデータをそろえる．反対に，ピッチポイントが 11 点以上存在するデータは，11 点目以降を切り捨てる．また，予測の場合 3 点目以降のピッチポイントは，シフト値が 10 cent を超えている場合のみ付与を行う．

なお，UTAU のポルタメントを形成する要素として，ピッチポイント間の線の形状を指定するパラメータが存在する．しかし，4.2 節の統計で約 95 %が曲線であることが判明している．そのため，本研究では線の形状は学習せずにすべて曲線を付与する．

6. 実験

本章では，5 章で提案したモデルの有用性を調べるために聴取実験を行う．聴取実験では，5 種類の音源を聴き比

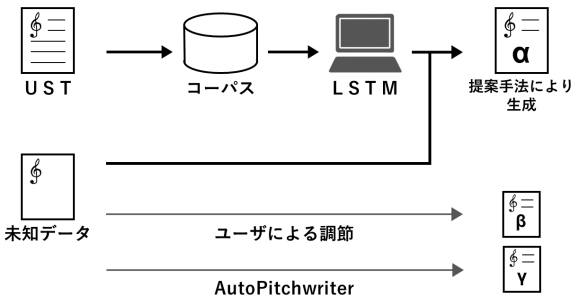


図 14　ヴィブラートのパラメータ推定 LSTM の実験手順

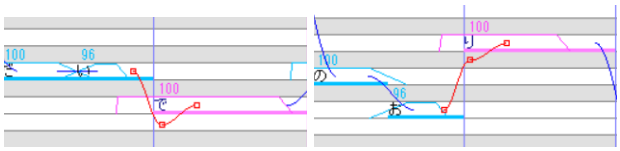


図 15　コーパス 2 の学習結果 (α 1)

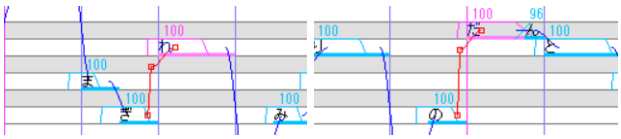


図 16　コーパス 2 の学習結果 (α 2)

べて，抑揚の有無と歌唱の自然さを評価する．

6.1　実験手順

実験の手順は，図 14 のとおりである．まず，収集した UST からコーパスの作成を行い，提案モデルで学習を行う．その学習結果を元に，未知の UST のピッチポイントの座標を予測し，ポルタメントの付与する．この過程により生成された UST を α とする．なお，本実験では出力データが 2 通り存在し，絶対的な表現の出力データで学習を行った結果を α 1，相対的な表現の出力データで学習を行った結果を α 2 とする．α とは別に，ユーザがパラメータ調節を行ったデータを β，従来手法である AutoPitchwriter で作成したデータを γ を用意する．聴取実験では，これらの音源にパラメータ未調節の音源を加えた全 5 種類の音声の比較を行う．

聴取実験前に 3 種類のコーパスを使用した実験を行う．1 つ目は，出現するポルタメントがすべて通常のポルタメントの楽曲 30 曲からなるコーパス 1 である．2 つ目は，複雑なポルタメントの割合が一定以上の楽曲 30 曲からなるコーパス 2 である．3 つ目は，収集した 1584 曲すべてを含むコーパス 3 である．

コーパス 1 は，α 1，α 2 共に全体的にデフォルト値に近い通常のポルタメントを付与することができた．コーパス 2 は，絶対的な表現を用いて学習した α 1 では，図 15 の左のようにしゃくりなどの歌唱表現を付与することができた．しかし，一部のサンプルでは，図 15 の右のように前の

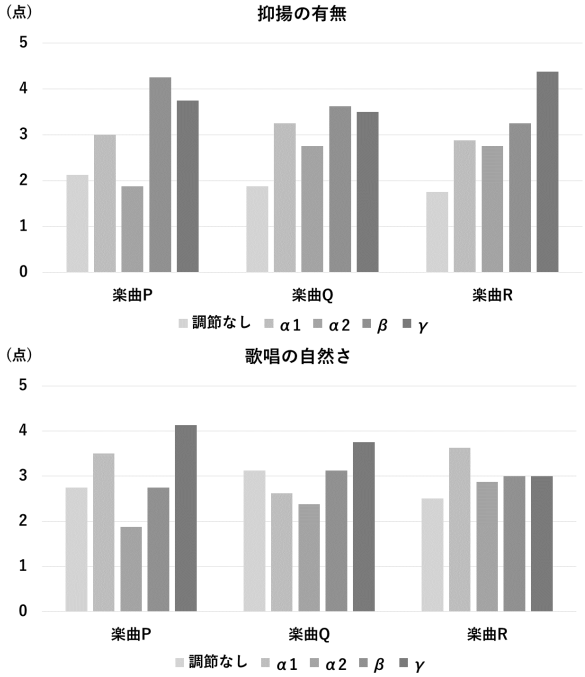


図 17　聴取実験の実験結果

音高との差よりもシフト値が小さく正しく歌唱表現が付与できていなかった．一方，相対的な表現を用いて学習した α 2 では，図 16 のようにピッチポイントの間隔が狭すぎて変化が急になる場合があった．コーパス 3 は，付与されるすべてのポルタメントが通常のポルタメントになった．

6.2　聴取実験

聴取実験では，6.1 節で解説した 5 種類の被験者に音声を聴き比べてもらう．なお，α 1 と α 2 はコーパス 2 の学習結果を元にポルタメントを付与している．被験者は 5 種類の音声を聴いて，抑揚の有無と歌唱の自然さ 2 つの項目についてそれぞれ順位付けする．順位付けで 1 位の音声は 5 点，2 位の音声は 4 点，3 位の音声は 3 点，4 位の音声は 2 点，5 位の音声は 1 点という点数に置き換えて統計を取る．

本実験では，計 8 名の被験者に音声を聴いてもらう．聴取する音声は P，Q，R の 3 曲分である．P は，BPM77 の楽曲 Q や R に比べて 4 分音符が多いスローテンポな楽曲である．R は，BPM154 で 8 分音符の多い楽曲である．R は，BPM172 で 16 分音符の多いアップテンポな楽曲である．なお，音声の生成に使用する音声ライブラリは，UTAU デフォルト音声 Ver1.2 である．これは，音声ライブラリに実験のノイズになりえる感情がこもった音声が含まれていないためである．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，5 種類の音源を聴く順番をランダムに入れ替えている．

聴取実験の結果は，図 17 のとおりである．α 1 と α 2 を比較した際，抑揚の有無と歌唱の自然さの両方の項目で α 1 の方が高い評価を得ている．しかし，α 1 の抑揚の有

無は，β や γ に比べて評価が低いという結果になっている．また，α 1 の歌唱の自然さは，楽曲 R では 5 種類の音声の中で最も評価が高い．一方，楽曲 Q では，パラメータが未調節の音源よりも評価が低い．

7. 考察

聴取実験前に，3 種類のコーパスを使用して学習を行った．出現するポルタメントがすべて通常のポルタメントの楽曲 30 曲からなるコーパス 1 では，想定される通常のポルタメントが正しく付与された．しかし，複雑なポルタメントの割合が一定以上の楽曲 30 曲からなるコーパス 2 では，一部のシフト値が前の音高差よりも小さくなったため抑揚が正しく表現できない場合があった．また，相対的な表現を使用した場合，ピッチポイントの間隔が狭すぎて抑揚が正しく表現できなかった．これはピッチポイントの少ないポルタメントは学習時に 0 埋めして長さを調節しており，その影響で 3 点目以降のピッチポイントのベクトルが想定よりも小さい値になったためだと考えられる．収集した 1584 曲すべてを含むコーパス 3 では，付与されるポルタメントがすべて通常のポルタメントになった．これは通常のポルタメントが複雑なポルタメントに比べて出現率が高いため，学習した際にすべて通常のポルタメントになったものだと考えられる．

聴取実験では，抑揚の有無と歌唱の自然さ 2 つの項目の評価を行った．この 2 つの項目を α 1 と α 2 に焦点を当てて比較した際，α 1 の方が高い評価を得た．これは，α 2 のピッチポイントの間隔が狭すぎて抑揚が正しく表現できず，歌唱が不自然なものになったためだと考えられる．また，α 1 を β や γ と比べた際，抑揚の有無の評価は β や γ の方が高くなった．これは，α 1 のシフト値が β や γ に比べて小さいためだと考えられる．また，一部の楽曲 Q では，抑揚のないパラメータが未調節の音源の方が α 1 よりも歌唱の自然さの評価が高くなった．このことより，抑揚の有無と歌唱の自然さは比例しないと言える．

8. おわりに

本研究では，ポルタメントの統計的性質の調査とポルタメントを学習するモデルの作成を行った．ポルタメントの統計的性質の調査では，ポルタメントの出現率とポルタメントの構成要素の分布について統計を取り，統計結果を元にモデルで学習するパラメータを決定した．ポルタメントを学習するモデルの作成では，音価，音高，音のタイプからそれぞれに付与されるポルタメントのピッチポイントの座標を予測するモデルを提案した．なお，出力データは，絶対的な表現と相対的な表現の 2 種類を用意し学習を行った．このモデルを用いて作成した音声を評価するために，聴取実験を行った．結果として，絶対的な表現と相対的な表現で学習した 2 種類の提案手法を比較した際，絶対的な表現の方が評価が高くなった．また，絶対的な表現で学習した提案手法と従来手法を比較した際，従来手法の方が評価が高くなった．

今後の課題は，大きく分けて 2 つある．1 つ目は，学習モデルや入出力データの改善である．本研究で提案した手法は，3 点目以降のピッチポイントが正しく学習できず，予測結果が想定と異なるものになった．今後は，モデルやデータの改善を行い，特殊な歌唱技法についても学習可能なものを提案する．2 つ目は，提案システムを異なる音声ライブラリで使用した際の比較である．先述のとおり，特殊な歌唱技法は音声データの少ない音声ライブラリから生成された合成音声の抑揚を改善するために誕生したものである．この歌唱技法を，音声データの多い音声ライブラリで使用した場合にどのような音声が生成されるか比較を行う．

参考文献

[1] 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）．
[2] 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）．
[3] 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）．
[4] 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）．
[5] SHACHI. NEUTRINO. 入手先 ⟨http://n3utrino.work/238/⟩, (2021.02.15).
[6] CeVIO Official Site. 入手先 ⟨https://cevio.jp/⟩, (2021.02.15).
[7] 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/⟩, (2021.02.15).
[8] Hochreiter, T and Jürgen, S.: LSTM can solve hard long time lag problems, In Advances in neural information processing systems, pp.473-479 (1997).
[9] 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）．