歌唱におけるタメ表現のモデル構築

藤田　千尋[*1]　竹川　佳成[*1]　平田　圭二[*1]

Model construction of representation to delay vocalization in singing

Chihiro Fujita[*1]　Yoshinari Takegawa[*1]　Keiji Hirata[*1]

Abstract – **In recent years, vocal music using singing voice synthesis system are increasing. But it is too difficult to sing a song similar to human for singing voice synthesis system. Furthermore, human tend to be a little off the point of vocalization on score. Specially, we took notice of delay vocalization in singing. We will construct a model of representation to delay vocalization in singing by Time-span tree and machine learning.**

Keywords : **Singing voice synthesis system, Time-span tree, Delay vocalization in singing**

1.　はじめに

近年，VOCALOID 等の歌声合成ソフトの普及により個人制作での歌声合成ソフトによる歌唱の入った楽曲が増加した．2015 年 7 月 9 日現在，動画投稿サイトであるニコニコ動画 [1] では VOCALOID カテゴリでの動画投稿数が 37 万件を超えており，同じく歌声合成ソフトである UTAU [2] のタグが付けられている動画数が 8 万件を超えているなど，歌声合成ソフトの普及が動画投稿サイトの隆盛にいかに貢献しているかがわかる．譜面に歌詞と音階等を打ち込むことによってその譜面通りの歌唱をしてくれるため，本人が歌唱を苦手としていても歌唱の入った楽曲の創作が以前よりも容易になったのが楽曲増加の要因の一つと言える．しかし，人間らしく自然に歌わせるためにはビブラートやしゃくり，グロウル等の細かい部分の調整を手作業で行う必要があり，多大な時間と労力を必要とする．また，人間の歌唱は必ずしも譜面のタイミングと同じではなく，少々前後にタイミングをずらして歌う傾向にある．特にその歌唱法についてはバラード，演歌などの比較的ゆっくりしたテンポの曲で顕著に出る傾向がある．

発声タイミングを楽譜上のタイミングよりも遅らせる歌唱表現を本論文ではタメ表現と呼ぶ．タメ表現は歌の旋律構造と歌に込められた意図に依存する．それにより，リスナーに感情や歌詞の強調を伝えることが可能となる．タメ表現を手動で歌声合成ソフトの歌唱に付与する場合には，制作者に音楽知識やリズム感が必要となる．

本研究の目的は，人間らしい歌唱表現としてタメ表現に焦点を絞り，歌声合成ソフトの歌唱に自動的にタメ表現を付与するためのタメ表現の分析とモデル構築である．

*1: 公立はこだて未来大学
*1: Hakodate Future University

2.　歌声合成ツール UTAU

歌声合成ツール UTAU [2] とは，飴屋／菖蒲 (あめや・あやめ) によって作成された Windows 向けの歌声合成ソフトウェアである．このツールではピアノロール形式で入力することにより，サンプリングされた音声ライブラリから歌唱を組み立てることが出来る．ソフトウェア本体に付属するデフォルト音源とは他に，ユーザが自身の声を録音し作成した音声ライブラリを使用することも可能であり，多様なユーザが作成した音声ライブラリが Web 上からダウンロード可能である．また，有志によって作成されたプラグインや合成エンジンにより，歌唱に効果を与えることや声色を変えることが可能である．

本研究では，UTAU を使用して UST (UTAU Sequence Text) ファイルを作成する．UST ファイルとは，UTAU によって作成される楽譜情報の入ったテキストファイルで，各音の音高や歌詞，音の長さ等の情報が入っている．音の長さの単位は Ticks で表される．本研究では UST ファイルの中の情報から主に音高と音の長さを使用する．UST 作成時に使用する音源はデフォルトのものとする．最終的に，UTAU のプラグインファイルとして UST ファイルにタメ表現を付与するプラグインの作成をする．

3.　関連研究

関連研究として，一般的な DTM に内蔵されている機能である乱数によって微妙にリズムをずらし人間が演奏しているように聞こえるようにするヒューマナイズ [3] がある．これは旋律に関係なく個々の音に乱数で数値を与えているため，人間らしさは多少付与でき

るもののその人間らしい動きに法則はない．

より人間に近い歌唱を生成するアプローチとして，人間的な要素を 1 つずつ取り入れるものと，人間の歌そのものを真似るものがある．前者のアプローチには，自然なビブラート [4]・ポルタメント [5]・グロウル [6] を生成する様々なモデルが提案されている．

後者のアプローチには，歌声合成ソフトに人間らしい歌唱をさせる VocaListener [7] がある．これは実際の歌唱データである音響信号をツールに入力すると，VOCALOID エンジンの合成パラメータの反復推定を行い入力された歌声に類似した歌声を合成する．入力した歌声に沿った音高やタイミング等のパラメータが出力される譜面に反映されるというものである．

4. 問題とアプローチ

本研究の課題は歌唱にタメ表現を付与するためのモデルを作成することである．また，歌唱にタメ表現を付与する場合にはどのような旋律にも対応できるように幅広く，かつ正確に付与できるようなシステムを作成する必要がある．その場合，モデル構築の際にどのような技術を使用するかが問題となる．これらの問題点を解決するために，以下の方法でモデル構築を行う．

4.1 楽曲分析結果としてのタイムスパン木

旋律と楽曲に含まれるタメ表現の構造を関連付けるために，音楽理論 GTTM (Generative Theory of Tonal Music) [8] のタイムスパン木を楽曲構造として採用する．GTTM とは，1983 年に Lerdahl と Jackendoff によって提案された音高の変動や拍節などから楽曲を簡約し，分析を行う音楽理論の一つである．タイムスパン木とは、GTTM の規則に基いて拍節構造やグルーピング構造からボトムアップに楽曲の構造を分析し、重要な音を主要部として二分木で図示される木構造である。楽曲のピッチイベントなどの譜面構造からタイムスパン木を作成し，構成規則に従って簡約化を行う．旋律中で隣接する 2 つの音に関して，より重要な音はタイムスパン木の幹枝 (primary) に割り当てられ，もう一方の音は副枝 (secondary) に割り当てられる．本研究では，最大タイムスパン (maximal time-span) [9] という概念を導入した．各々のピッチイベントの情報にタメ表現による発声タイミングのズレの重み付けを行うことにより，各旋律にどれだけのタメが生じるかを分析する．これらの分析結果をタイムスパン木作成の際に用いる．それにより，簡約化されたタイムスパン木をコーパスに関連付けることによって旋律構造が関連付けされたコーパスを作成することが出来る．

図 1 にタイムスパン木，最大タイムスパン階層，タメ時間の関係を図示する．サンプル曲は平井堅「大きな古時計」4:11 秒から 4:17 秒までの部分であり，(c) にその楽曲部分の五線譜表現を示す．(a) はその楽曲部分の分析結果であるタイムスパン木を，(b) は最大タイムスパン階層を，(d) は楽曲部分に含まれる各音のタメ時間を表す．ここで (a) より，構造的に最も重要な音は「そのとけい」の「い」の音である．(b) では，各最大タイムスパンに関連付けられたタメ時間を ticks 単位で示す．例えば 2 番目の音である D♭4 の「の」の音に関して，3 つの最大タイムスパンが重なっているということは，タイムスパン木において root から分岐 3 つ経たレベルに出現していることを意味する．「の」のタメ時間は，それより上位（タイムスパン木の根に近い）のタメ時間の和として定義され，30 + (-30) - 60 = - 60 となる．(d) において各音のタメ時間が正の時は，棒グラフが中央より上に伸び，負の時は下に伸びる．「そ」と「け」の Y 軸の値は 0 であり，つまりタメて歌われておらず，楽譜通りのタイミングで発声しているという意味である (deadpan)．

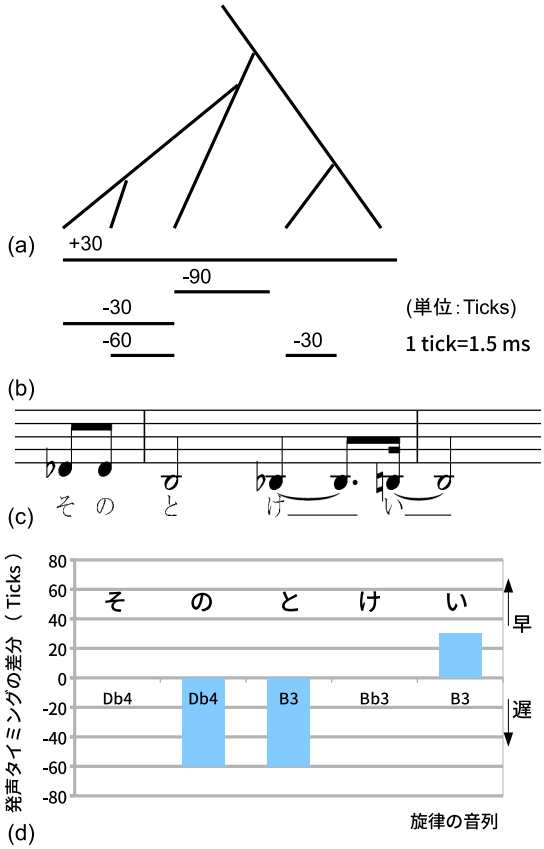

図 1 タイムスパン木に基づくタメ時間の生成メカニズム

Fig. 1 Formation mechanism of delay vocalization in singing based on Time-span tree

4.2 機械学習によるタメ表現のモデル構築

上記のタイムスパン木により旋律構造を関連付けられたコーパスから，タメ表現のモデルを構築する．複数のコーパスから旋律構造とタメ表現のタイミングを関連付ける統計モデルを構築する．旋律構造とそのタメ表現上での発声タイミングを関連付ける機械学習器を構成し，作成したコーパスを用いて学習させることによってモデル構築を行う．

5. これまでの成果

5.1 タメ表現コーパス作成

これまでの主たる成果は，RWC 研究用音楽データベース [10]RWC-MDB-PR-2001 内の 22 曲，その他の曲 10 曲の合計 32 曲からなるタメ表現コーパスである．以下の手順で作成する．まず歌唱に含まれているタメ表現のサンプル部分集め，そのサンプル部分を UTAU を使用して打ち込み，タメ表現の含まれているものと含まれていないものを比較することによりタメ表現のコーパス 1 曲分を作成する．

サンプル集めの方法として，RWC 研究用音楽データベース RWC-MDB-PR-2001 に収録されているものと RWC の Web ページからダウンロードできる MIDI 音源中の歌唱部分に当たるパートを聴き比べ，MIDI 音源の方をタメ表現の含まれない楽譜通りの歌唱，実際に歌唱されている方をタメ表現の含まれている歌唱と仮定した．次にその歌唱の中で明らかに発声タイミングのずれている箇所を ust ファイルの譜面上に書き起こし，それら 2 つのファイルを比較することによってコーパスを作成した．7 月 17 日現在，RWC 研究用音楽データベース RWC-MDB-PR-2001 内から 22 曲，それらの曲中から 30 箇所程度のタメ表現を含んでいると判断した部分のコーパスを作成した．

図 2 に，コーパス作成に使用する楽譜通りの UST ファイルと実際の歌唱通りの UST ファイルを図示する．使用した楽曲は図 1 で使用したものと同楽曲である． 図 3 は，図 2 での UST ファイルを WAV ファイルとして出力し，波形化した図である．最上部が実際の楽譜であり，上の波形図が楽譜通りに打ち込みをして出力した WAV ファイルの波形，下の波形図が実際の歌唱通りに打ち込みをして出力をした WAV ファイルの波形である． 図 3 を見ると，タメ表現の含まれる歌唱部分では楽譜通りの歌唱に対して実際の歌唱通りの方が発声のタイミングが遅れているのが明らかである．

5.2 UST 用タイミング比較ツール

コーパス作成に使用した楽譜通りの歌唱を打ち込みをした UST ファイルと実際の歌唱通りに打ち込みをした UST ファイルの 2 つを入力とし，それぞれのファイルから音高，音の長さの情報を取り出して音の長さの情報からそれぞれの発声タイミングを計算し，楽譜通りの発声タイミングと歌唱通りの発声タイミングの差を CSV ファイルに出力するツールを Java で作成した．出力される値は楽譜通りの歌唱の発声タイミングと実際の歌唱通りの発声タイミングの差分の値である．単位は Ticks である．表 1 に現時点で実際に入力した UST ファイル中の各音符の長さを，表 2 に出力された CSV ファイルに記載された表の一部分を示す．

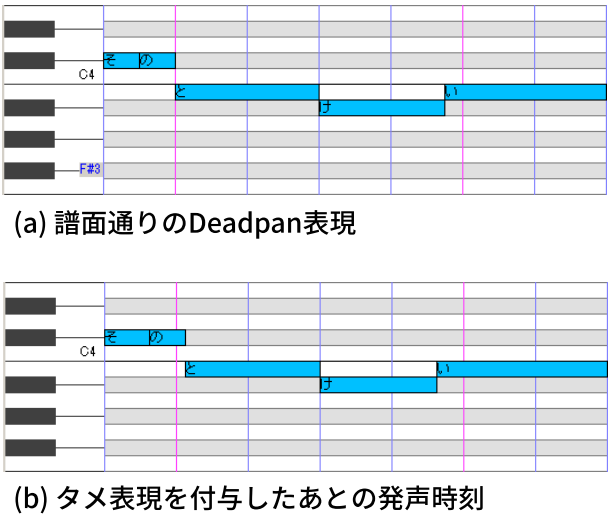

図 2 タメ表現によって変化する発声時刻
Fig. 2 Change the time of vocalization by delay

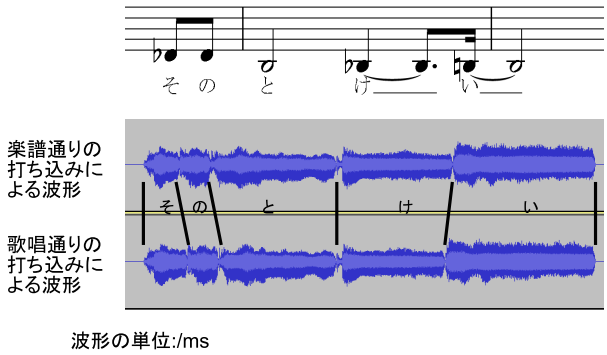

図 3 楽譜通りの打ち込みの波形と歌唱通りの打ち込みの波形
Fig. 3 waveform according to score and waveform according to singing

表 1 入力した UST の各音符の長さ
Table 1 The length of each notes at imported UST

音高	Db4	Db4	B3	Bb3	B3
楽譜通り	240	240	960	840	1080
歌唱通り	300	240	900	780	1140

表 2　UST 用タイミング比較ツールの出力
Table 2　Export tool of compared timing used UST

音高	Db4	**Db4**	**B3**	**Bb3**	**B3**
タメ	0	-60	-60	0	30

6.　終わりに

本研究では，歌声合成ソフトによって入力された歌唱へのタメ表現付与のために旋律構造と関連付けたタメ表現モデル作成することを目的とした．今後の予定として，サンプル数を増やした上で音楽理論 GTTM に基づいたタイムスパン木を用いてのコーパス作成を行い，モデル構築を予定している．そのモデルを構築し，評価実験を行った上で作成した UST ファイルの歌唱にタメ表現を付与するシステムの作成を行う．

また，モデルの評価方法として cross validation 法，もしくは関連研究で挙げた VocaListener での評価基準である歌詞アラインメントの誤り訂正機能の有効性，パラメータの反復推定の有効性，音源データの違いに対する頑健性を参考に，本評価を行う予定である．

7.　謝辞

本原稿を作成するにあたり，文章のチェックやアドバイスを下さった平田・竹川研究室の皆様に感謝致します．

参考文献

[1] ニコニコ動画; http://www.nicovideo.jp/(2015 年 7 月参照).
[2] 飴屋／菖蒲: 歌声合成ツール UTAU　サポートページ; http://utau2008.web.fc2.com/(2015 年 7 月参照).
[3] 奥平, 平田, 片寄:ドラム演奏の打点時刻及び音量とグルーブ感の関連ポップス系ドラム演奏の打点時刻及び音量とグルーブ感の関連について (第 3 報)– データの基礎的分析とドラム演奏生成システムの実装 –; 情報処理学会研究報告音楽情報科学, 2006-MUS-064, pp53-58(2006).
[4] 右田, 森勢, 西浦:歌唱データベースを用いたヴィブラートの個人性の制御に有効な特徴量の検討; 情報処理学会論文誌, Vol.52, No.5, pp1910-1922(2011).
[5] 辰巳, 森勢, 片寄:歌唱特徴付与システム「ロックボーカルレゾネータ」; 情報処理学会研究報告, Vol.2010-MUS-87, No.7(2010).
[6] Jordi, Merlijn, 他:スペクトルモーフィングによるグロウル系統の歌唱音声合成; 情報処理学会研究報告, Vol.2013-MUS-100, No.24(2013).
[7] 中野, 後藤: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, pp3853-3867(2011).
[8] Lerdahl, Jackendoff:A Generative Theory of Tonal Music, The MIT Press(1983).
[9] Tojo, Hirata:Distance and Similarity of Time-span Trees; Journal of information processing, Vol.21, No.2, pp256-263(2013).
[10] 後藤, 橋口, 他:RWC 研究用音楽データベース : ポピュラー音楽データベースと著作権切れ音楽データベース; 情報処理学会研究報告音楽情報科学, Vol. 情報処理学会研究報告音楽情報科学, pp35-42(2001).