3B1-OS22c-10

階層 Pitman-Yor 言語モデルを用いた自動作曲手法の提案

A proposal of the automatic musical composition method using hierarchical pitman-yor language model

白井 亨*1　　谷口 忠大*1
Akira Shirai　Tadahiro Taniguchi

*1立命館大学
Ritsumeikan University

Recently, a lot of researchers in the field of an automatic musical composition use a n-gram model to generate music. An order n is a very important variable. A high order model can generate a melody which is well featured by training data. However, possible states exponentially explode when the n increases. Furthermore, which causes a lack of training data. Thus, we propose a method using Variable-order Pitman-Yor Language Model, extension of Hierarchical Pitman-Yor Language Model, which generates new melodies based on variable orders with appropriate smoothing. We also evaluate this method by some experiments.

1. はじめに

近年，パソコン上で作曲を行う DTM が広く普及したことによって楽器などを持っていなくても手軽に作曲環境を手に入れることができるようになった．しかし現在，作曲は音楽家，音楽学校で専門的な指導を受けた人もしくは趣味として音楽を勉強した人が行うのが一般的である．音楽的な知識を持ち合わせていない人たちにとっては前述の人たちが作曲した楽曲を聴き，楽しみ，共有しあうことが今日の日常的な音楽スタイルであり作曲を行うことはないだろう．なぜなら作曲を行うためには様々な音楽的知識が必要であり，これらを持っていない音楽初心者にとっては作曲は未だに気軽に取り組めるものではないからである．よって，そのような人たちが作曲を行うためには音楽的な補助が必要になる．そこで，本研究では音楽的な知識を有しないユーザでも容易にイメージ通りの作曲ができるように補助を行うための手法を提案する．

2. 研究目的

音楽的な知識を持たない人でも手軽に作曲を行うことのできるポピュラーな方法として自動作曲がある．自動作曲システムはフリーソフト*1としても多く広まっておりユーザが楽曲イメージ（ジャズ風，明るいなど）を選ぶことでそのイメージに合う楽曲を生成することが出来る．しかし，一般的な自動作曲システムでは楽曲イメージはあらかじめ用意されたものに限られ，ユーザの望むイメージがなかった場合や生成された楽曲がユーザの思っていたイメージと違っていた場合，望み通りの作曲を行えないという問題がある．このような問題に対して対話型進化計算を用いユーザの感性をシステムに取り入れて対話的にユーザの曲想に合う楽曲を生成する研究が行われている [高木 98]．この方法を用いることでユーザは自分の曲想に合う楽曲を生成することができると言われる．しかし，このような計算法を用いた作曲手法では解として数十秒から数分あるメロディーをすべて聞いて評価することになりユーザの疲労が増大してしまうため，結果的に世代数を少なくする必要があり適切な解に近づけないという問題がある．実際このタイプの自動作曲システムは一般に普及していない．

一方，最近メディアでもよく取り上げられている自動作曲システムとして [深山 08] らのオルフェウスがある．オルフェウスは入力された日本語の韻律を制約として動的計画法により自動作曲を行う．また，システムは Web 上で公開されており，反響を呼んでいる．歌詞やコード進行を設定できる自由度や Web 上で作曲が可能な手軽さがオルフェウスの強みであると考えられるが，一方で最適化手法を用いていることで多様性を失っているという問題もある．加えて，生成結果に対して操作を行うことができない点や，計算の過程ですでに確定した部分に影響を及ぼす制約式を加えることで式が複雑化していき解の探索が困難になるなどの問題点がある．

以上の問題を解決する方法として，[白井 10] では統計的言語モデルとしてよく用いられる n-gram モデルから確率的にメロディーを生成する手法を提案している．n-gram の n を 2 としてモデルの学習を行い，そこからギブスサンプリングを用いてメロディーの生成を行うことで，多様なメロディーを生成することが可能である．また，コード進行による重み付けや，楽曲構造による制約，部分的な修正などの操作を行うことでユーザの評価を得られることが示されている．しかし，n-gram 長が 2 と短いため学習データから特徴抽出する機能に乏しく，楽曲構造による制約を用いて既存楽曲に似ているメロディーの生成を試みているが，良い評価は得られていない．よって，学習データの特徴を持つメロディーを生成するためにはオーダー n を大きくする必要がある．しかし，n が大すぎると学習データの再現率は上がるが汎化能力は下がってしまう．また，n が増えると状態数が指数的に爆発するため学習データが不足するという問題もある．よって，学習データから効率よく特徴を抽出するためには文脈によって適切な n グラム長を推定しながら，スムージングを行うモデルが必要である．

そこで本研究では Variable-order Pitman-Yor Language Model(VPYLM) を用いてメロディーのモデル化を行う．また本研究では自然なメロディーを生成するために学習データからの特徴のみではなく，得られたモデルに対してコード進行と歌詞を与えることで適切なメロディーをサンプリングする手法を提案する．

連絡先: 白井亨，立命館大学理工学研究科，創発システム研究室，
滋賀県草津市野路東 1-1-1，shirai@em.ci.ritsumei.ac.jp

*1　音楽研究所: 自動作曲システム ACS,
http://hp.vector.co.jp/authors/VA014815/music
鵜飼利彦: Juice and Candy 3.33,
http://www.vector.co.jp/soft/win95/art

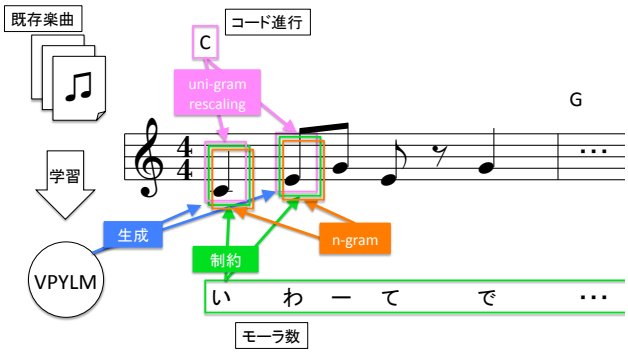

図 1: 提案手法の概要図

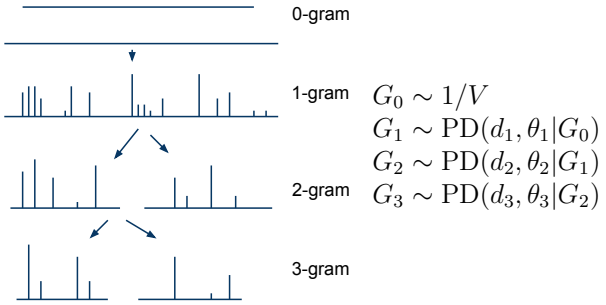

図 2: 階層的な n-gram 分布の生成

3. 提案手法

本研究ではユーザが与える歌詞とコード進行に対して学習データ中に含まれる既存楽曲の特徴を持ったメロディーを生成することを目的とする．提案手法の概要を図 1 に示す．

3.1 楽曲コーパスの作成

既存楽曲を学習データとして用いるために楽曲コーパスを作成する．メロディーを音符の列 $\mathbf{s} = (s_1, s_2, ..., s_N)$ として N 次元のベクトルで表す．ベクトルの各要素はメロディーを離散的に分割したものであり本研究では 8 分音符単位で分割を行う．よって，ベクトルの各要素には音符のピッチ，音が伸びている状態，休符のいずれかが入ることになる．また，同様にしてコード進行を $\mathbf{c} = (c_1, c_2, ...c_N)$ で表す．こちらの各要素にはコードネームが入る．なお，調が C でない楽曲については C 調に移調を行う．

3.2 VPYLM の学習

n-gram モデルでは n が小さいと学習データの再現率が下がり，逆に大きいと状態数が指数的に爆発してしまうため学習する音符によって潜在的な n-gram 長を推定する必要がある．また，学習データ量の不足によって本来複数存在する可能性のある状態がほとんど現れない場合があるため，適切にスムージングを行う必要がある．そこで本研究では Hierarchical Pitman-Yor Language Model(HPYLM) を拡張した VPYLM を用いてメロディーの学習を行う．HPYLM とは Teh らによって提案された n-gram 分布の階層的な生成モデルである（図 2）[Teh 06]．HPYLM では Hierarchical Pitman-Yor Process と呼ばれるノンパラメトリックな確率過程によって，適切にスムージングされた n-gram 分布を階層的に生成及び推定することが可能である．また，現在数あるスムージング手法の中でも最高性能と言われる Kneser-Ney スムージング [Kneser 95] はこの確率過程の近似となっている．HPYLM では学習を行う際に確率的に接尾辞木の一つ上の文脈をカウントすることでスムージングを行う．

本手法ではこの HPYLM を拡張したモデルである VPYLM を用いる．VPYLM は持橋らによって提案された可変長 n-gram 言語モデルである [持橋 07]．このモデルでは n-gram 長を固定せず，確率的に接尾辞木のノードを辿ることで単語の持つ潜在的な n-gram 長を推定することが可能である．

3.3 メロディーの生成

本手法では歌詞とコード進行を与えた上でメロディーの生成を行う．具体的には歌詞の文字数による制約とコード進行による制約をかけながらサンプリングを行う．

3.3.1 Gibbs sampler によるメロディーの生成

ある文脈 $\mathbf{h}$（ここでは音符の並び）の後に音符 s が出現する確率は，n-gram 長 n を隠れ変数とみなして

$$\begin{aligned} p(s|\mathbf{h}) &= \textstyle\sum_n p(s, n|\mathbf{h}) \\ &= \textstyle\sum_n p(s|n, \mathbf{h}) p(n|\mathbf{h}) \end{aligned} \tag{1}$$

のように予測を行う．ここで，第一項はオーダーを n とした HPYLM の予測確率，第二項は文脈 $\mathbf{h}$ の持つ n-gram 文脈長分布である．式 (1) より，メロディー全体の事後確率（メロディーの生成確率）は

$$p(\mathbf{s}) = \sum_i^N p(s_i|\mathbf{h}) \tag{2}$$

のように求めることができる．ギブスサンプリングするためには生成確率の比が求まればよいので式 (2) から音符 s_i 以外の音符 $\mathbf{s}_{-i}$ が決まっているとに s_i が音符 k になる確率は

$$\begin{aligned} p(s_i = k|\mathbf{s}_{-i}) \quad &\propto \quad p(s_i = k|\mathbf{h}) p(s_{i+1}|\mathbf{h}') \\ &\times \quad ... \times p(s_{i+n_{\max}}|\mathbf{h}^{(n_{\max})}) \end{aligned} \tag{3}$$

で求めることができる．ここで $n_{\max}$ は学習する際に決めた VPYLM の最大 n-gram 長である．式 (3) を用いることでメロディーの Gibbs Sampler を構成することができるが，本手法ではさらに歌詞とコード進行によって重み付けを行う．

3.3.2 歌詞の割り当て

歌詞の総モーラ数[*2]を M，メロディーに含まれる伸ばしている状態，休符以外の総音符数を N_{note} とする．ここではモーラ数は音符数を上回ることはないとし，共に同じ数存在する状態を理想型と考える[*3]．よって M が N_{note} より大きくなると急激に確率が下がり，同じ時に最大値をとり，M が小さくなると徐々に確率が下がるような制約を導入する．本手法ではこのような制約を以下の式で表す．

$$p(M, N_{\text{note}}) = \begin{cases} \exp((N_{\text{note}} - M)\alpha) & \text{if } N_{\text{note}} < M \\ \exp(\frac{M - N_{\text{note}}}{\beta}) & \text{otherwise} \end{cases} \tag{4}$$

ここで α 及び β は指数関数の減衰率を調整するパラメータである．3 にこの制約の振る舞いを示す．式 (3) に式 (4) を適用

*2 一定の時間的長さをもった音の文節単位．音節とは異なり長音，促音，撥音もモーラ数に含まれる．

*3 モーラ数が音符数より小さい場合，モーラを発音している間にピッチが変わることを許す．

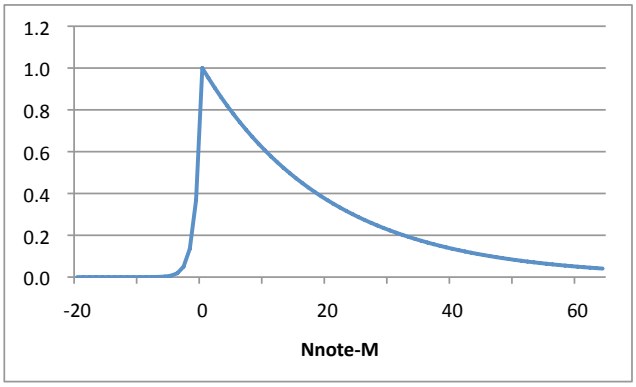

図 3: 制約の振る舞い（$\alpha = 1, \beta = 20$）

すると

$$\hat{p}(s_i = k|\mathbf{s}_{-i}) = p(s_i = k|\mathbf{s}_{-i})p(M, N_{\text{note}}) \quad (5)$$

となり，歌詞を考慮したメロディーの Gibbs Sampler を構成することができる．

3.3.3 uni-gram rescaling を用いたコード進行のトピック適応

コード進行による制約をトピック適応の問題と捉え uni-gram rescaling を用いて適応を行う．コードを適応すると

$$p(s_i|s_{i-1}^{i-n}, c_i) \propto \frac{p(s_i|s_{i-1}^{i-n})p(s_i|c_i)}{p(s_i)} \quad (6)$$

のように表すことができる．ここで $p(s_i|c_i)$ はコード c_i 上で音符 s_i が出現する確率であり，楽曲コーパスから学習する．式 (5) に式 (6) を適用すると

$$\begin{aligned} p(s_i|\mathbf{s}_{-i}, \mathbf{c}) \quad &\propto \quad p(s_i = k|\mathbf{h}, c_i)p(s_{i+1}|\mathbf{h}', c_{i+1}) \\ &\times \quad \ldots \times p(s_{i+n_{\max}}|\mathbf{h}^{(n_{\max})}, c_{i+n_{\max}}) \\ &\times \quad p(M, N_{\text{note}}) \end{aligned} \quad (7)$$

となる．ここで $p(s_i|\mathbf{h}, c_i)$ は

$$p(s_i|\mathbf{h}, c_i) = \frac{\sum_n p(s_i|n, \mathbf{h})p(n|\mathbf{h})p(s_i|c_i)}{p(s_i)} \quad (8)$$

と表させる．式 (7) を用いて Gibbs Sampler を構成することで歌詞及びコード進行を考慮したメロディーを生成する．

4. 評価実験

提案手法の妥当性を検証するために 20〜30 代の男性被験者 5 名に対して評価実験を行った．

4.1 実験条件

以下の条件で実験を行った．

学習データ 楽曲コーパスとして日本の Pops 曲から A，B メロ部分を 35 曲使用した．コーパス量が少ないため対策として同じデータを 10 個分学習させた．総単語数は 37760，総語彙数は 30 となった．

学習設定 1000 回のギブスサンプリングを行い，n-gram の最大長 $n_{\max}$ を 10 とした．

生成設定 VPYLM から 100 回の burn-in の後 200 回のギブスサンプリングを行い最大生成確率のメロディーを生成した．メロディーの長さ N を 64（8 小節）とした．また歌詞の制約パラメータを $\alpha = 10, \beta = 50$ に設定した．よって，生成されるメロディーの音符数は歌詞のモーラ数より多くなりやすくなっている．本手法では音符数が歌詞のモーラ数より多い場合の割り当て方が決まっていないため，手動で割り当てを行った．

比較対象 本手法の性能を評価するため，2 章で上げた 2-gram を用いる手法及びオルフェウスからそれぞれ同じコード進行，歌詞の条件の下で生成を行った．

評価項目 それぞれの手法で生成したメロディーを被験者に聴き比べてもらい，3 つの項目ごとに順位をつけてもらった．評価項目はそれぞれ，メロディーが良い，歌詞とメロディーが合っている，ベタなメロディーである．3 手法× 2 セットで合計 6 曲を用いて実験を行った．

4.2 実験結果

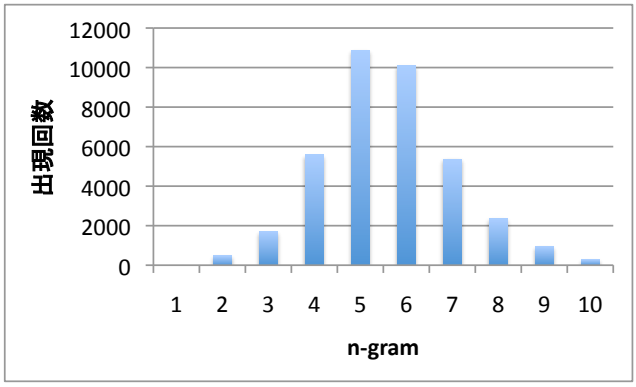

図 4: n-gram 長の分布

図 4 は学習した VPYLM の n-gram 長分布である．5, 6-gram 文脈になる単語が多く，比較的長い系列を扱えることがわかる．

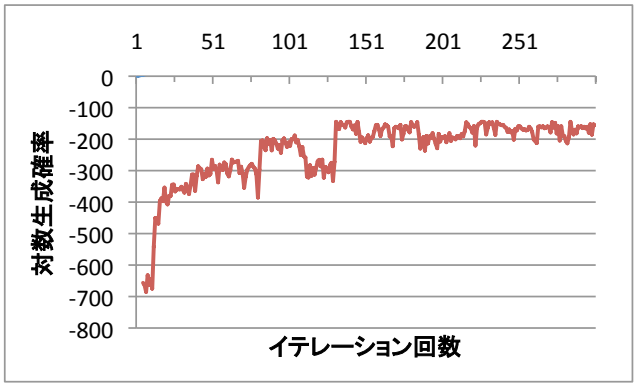

図 5: 生成確率の推移

図 5 は VPYLM からメロディーを生成する際に生成確率がどのように変化するかを表している．縦軸は生成確率の対数をとったもの，横軸はサンプリング回数を表している．最適化ではないため値は常に変動するが徐々に事後確率の高いサンプルが生成されていることがわかる．

図 6 はそれぞれの評価項目について各手法の順位を合計したものである．オルフェウスはメロディーの良さと歌詞とメロ

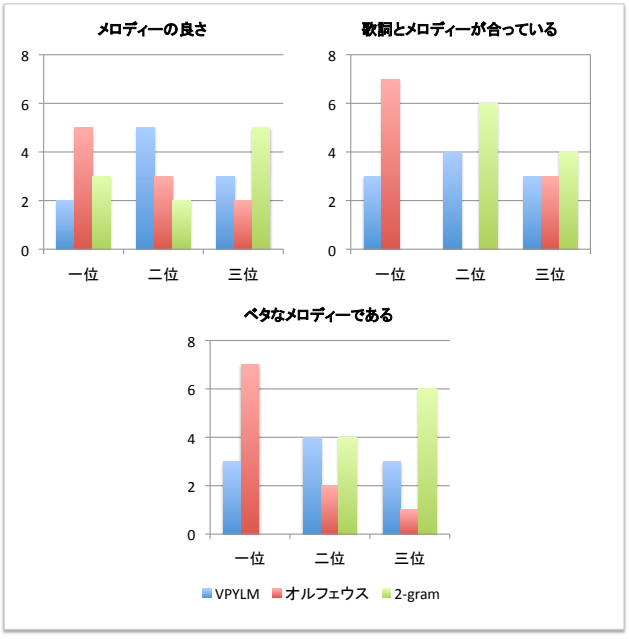

図 6: 評価数の比較

ディーが合っているが共に一位の合計が高いが，ベタさでも一位が高くなっている．2-gram を用いる手法は逆に三位が高くなっている．一方本手法はどの項目においても平均的に評価を得ている．

5. 考察

本稿では HPYLM を可変長 n-gram モデルに拡張した VPYLM を用いた自動作曲手法を提案し，有効性について検証を行った．実験結果より学習した VPYLM よりサンプリングを行うことで事後確率の高いメロディーを得ることが可能であることが実証された．また，HPYLM による n-gram 分布の階層的なスムージングにより従来手法で存在したコーパス量不足の問題点を緩和したと言える．しかし，本稿における本手法と他手法との比較では目立って有効な点は見られなかった．特にオルフェウスとの差異が顕著であった．これは，オルフェウスで考慮されているリズム及びこれと歌詞の関係が本手法では考慮されていない点などが原因だと考えられる．実際，歌詞を全く考慮していない 2-gram を用いる手法は歌詞に対する評価が低いという結果が出ている．よって，リズムに関する制約を加えることが今後の課題になる．

また，今回の実験では触れなかったが，VPYLM を使用する利点の中に長い文脈を学習できる点がある．よって，学習データとして使用するコーパスの特徴を持つメロディーを生成可能であると考えられる．さらに意図的にコーパスを操作することで特定の楽曲成分を強くした生成も可能だと思われる．これらについては今後検証する必要がある．

今後，VPYLM を基に本手法を拡張していくことで，よりユーザの満足を得られる自動作曲システムに近づけたい．

参考文献

[白井 10] 白井亨，谷口忠大: ギブスサンプリングを用いたインタラクティブ作曲システムの提案，ヒューマンインタフェースシンポジウム 2010，論文集，3412 (2010).

[高木 98] 高木英行，畝見達夫，寺野隆雄: 対話型進化計算法の研究動向，人工知能学会誌，Vol.13, No.5, pp.692-703 (1998).

[深山 08] 深山覚，中妻啓，米林裕一郎，酒向慎司，西本卓也，小野順貴，嵯峨山茂樹: Orpheus:歌詞の韻律に基づいた自動作曲システム，情報処理学会研究報告，2008-MUS-76, pp.179-184，2008.

[Teh 06] Teh, Y.W.: A Hierarchical Bayesian Language Model based on Pitman-Yor Processes, Proc.COLING/ACL 2006, pp.985-992 (2006).

[Kneser 95] Kneser, R. and Ney, H.: Improved backing-off for m-gram language modeling, Proc. ICASSP, Vol.1, pp.181-184 (1995).

[持橋 07] 持橋大地，隅田英一郎: 階層 Pitman-Yor 過程に基づく可変長 n-gram 言語モデル，情報処理学会論文誌，Vol.48, No.12, pp4023-4032 (2007).