

2次元LRパーサに基づく実演奏MIDI信号からの自動採譜

高宗 典玄^{1,a)} 亀岡 弘和^{1,b)} 嵯峨山 茂樹^{2,c)}

概要: 自動採譜は音楽音響信号から自動でその楽譜を推定する問題であり、大きく分けて、音響信号からの楽器がいつどの高さの音を鳴らしたかを推定する多重音解析と多重音解析で得られた結果から実際の楽譜を推定するリズム解析の2つがある。本研究では、後者のリズム解析に焦点をあて、MIDI信号から譜面情報の推定する問題を扱う。リズムとテンポの不確実性があるこの問題に対し、それぞれの生成過程を考え、統合的に解くことによりその解決を試みる。ここで、リズムの生成モデルとして2次元PCFGを用い、その解析アルゴリズムとして2次元LRパーサを提案する。実際に人間が演奏したいいくつかのMIDI信号に対し採譜実験を行い、その結果を示す。

キーワード: 自動採譜, リズム解析, MIDI信号, 2次元LRパーサ

Automatic Transcription from MIDI Signals of Music Performance Using 2-Dimensional LR Parser

Abstract: This paper proposes 2-dimensional LR parser for automatic music transcription of polyphonic MIDI signals. Automatic music transcription are roughly divided into two subproblems that are interdependent of each other: rhythm recognition and tempo estimation. This paper proposes to model the generative process of music performance by combining the sub-process by which a set of onset times and offset times in music score is generated and the sub-process by which a tempo curve is generated, and develop the analysis algorithm for this generative model. We show the transcription results obtained with the present method.

Keywords: automatic transcription, rhythm recognition, MIDI signals, 2-dimensional LR parser

1. はじめに

自動採譜とは、人間が演奏した音楽音響信号から自動で楽譜を推定することであり、音声信号処理における音声認識の位置に対応する、音楽音響信号処理における重要な課題のひとつである。

自動採譜には大きく分けて2つの段階がある。1つは音響信号から演奏している楽器やそれぞれの音の音高や発音時刻、消音時刻を推定する多重音解析であり [1], もう1つは多重音解析で推定された音高や発音時刻、消音時刻からテンポや音価を推定するリズム解析である [2], [3], [4]。本

研究では、後者のリズム解析の問題を取り扱うため、人間が演奏したMIDI信号からの自動採譜を目指す。

リズム解析の問題は、人間の演奏が多くの場合テンポ変動や発音・消音時刻のゆらぎを含んでいるため、非常に困難な問題となっている。これは、観測上の時間においての音の長さは、楽譜上の音の長さやテンポに依存するため、どのような楽譜から演奏されたかやどのようなテンポ変動を行ったかを一意に分解することは不可能であることに由来する。しかし、我々は多くの場合、そのようなテンポ変動や発音・消音時刻のゆらぎを含んだ演奏を聴いても、楽譜やテンポを推定することが出来る。これは、人間は楽譜やテンポに対して先見知識があり、その先見知識に沿うような楽譜とテンポを推定していると考えられる。そこで、本研究では先見知識として楽譜やテンポの確率的生成モデルを構築し、それを確率的逆問題として楽譜やテンポを同時に推定することを考える。

本稿では、2章で生成モデルの説明を行い、その解析ア

¹ 東京大学情報理工学系研究科
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan

² 明治大学総合数理学部
Meiji University
4-21-1 Nakano, Nakano-ku, Tokyo, 164-8525, Japan

a) takamune@hil.t.u-tokyo.ac.jp

b) kameoka@hil.t.u-tokyo.ac.jp

c) sagayama@meiji.ac.jp

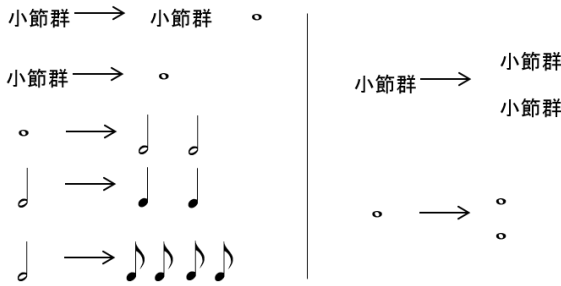


図 1 2次元 PCFG の文法例．左側は時間方向の文法，右側は音高方向の文法の例になっている．

ルゴリズムを 3 章で行う．そして，提案手法の動作を確認するために採譜の実験を 4 章で行い，本稿のまとめを 5 章で行う．

2. 音楽演奏 MIDI 信号の生成モデル

2.1 楽譜の生成モデル

音楽を時間方向に見ていくと，モチーフやフレーズなどといった階層構造を持つことが分かる．これは，自然言語が単語や文節，文といった階層構造を持つことと似ているため，そのような階層構造を扱う自然言語処理で用いられているモデルである確率文脈自由文法 (Probabilistic Context-Free Grammar; PCFG) を利用できないかということが考えられる．一方，音楽には声部やパートといった，音高方向の階層構造も有しているため，単純に PCFG を音楽に適用することは出来ない．そこで，このような，時間方向と音高方向の二つの方向の階層構造を表すモデルとして，我々の研究室で提案された 2 次元 PCFG[5] を用いることを考える．2 次元 PCFG の生成規則の例を Fig. 1 に示す．これを用いることで，起こりうるすべての楽譜に対し，少ない生成規則で生成確率を計算することができるようになる．ここで実際の楽譜で頻出するリズムが出現する確率を高く，あまり現れないリズムが出現する確率を低く設定することで，楽譜の尤もらしさを生成確率で表現できる．

2.2 テンポ・発音時刻・消音時刻の生成モデル

人間が音楽を演奏する際，テンポは必ずしも一定ではなく，緩やかな変動を伴っている場合がほとんどである．また，各音の発音時刻や消音時刻も物理的な制約や，演奏上の表情付けで楽譜上の位置からずれることがある．そこで，Fig. 2 に示すような有限状態空間を用いてテンポ変動や発音時刻・消音時刻のずれの生成をモデル化することを考える．これは，各状態が楽譜上のどの位置にいるのかを表し，演奏の時間がある単位時間進むごとに状態を遷移し，楽譜上の発音時刻，消音時刻に対応する状態の前後の状態から発音や消音の指令を出力するモデルである．ここで，

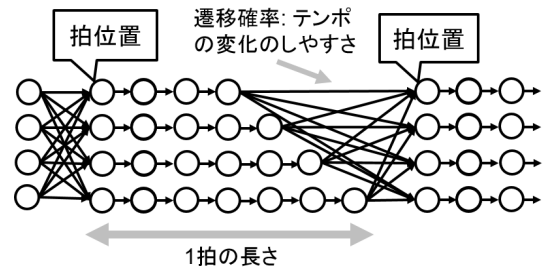


図 2 テンポ・発音時刻・消音時刻を生成する有限状態空間



図 3 階層構造の解釈の不確定性を示す例

各 left-to-right のパスの中の状態数とそのパスを通るときの 1 拍の時間を表し，パス間の遷移確率は 1 拍の長さがどう変化しやすいか，つまりテンポ変動の起きやすさを表す．

そこで，ここではテンポ変動の滑らかさを表現するために，拍の長さが v の拍から拍の長さが τ の拍への遷移確率 $\phi_{\tau,v}$ を

$$\phi_{\tau,v} \propto \mathcal{N}(\tau; v, (C^{\text{trans}} v)^2) \quad (1)$$

と設計する．ただし， $\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ であり， C^{trans} はテンポの変化を表す定数である．また，発音時刻，消音時刻のずれを表現するために，楽譜上の発音時刻又は消音時刻に対応する状態 l に対して，状態 k がその発音時刻又は消音時刻を生成する確率を状態間の距離 $d_{k,l}$ を用いて

$$p_{k,l}^{\text{onset}} \propto \mathcal{N}(d_{k,l}; 0, (C^{\text{onset}} \tau)^2), \quad (2)$$

$$p_{k,l}^{\text{offset}} \propto \mathcal{N}(d_{k,l}; 0, (C^{\text{offset}} \tau)^2) \quad (3)$$

と設計する．ただし， C^{onset} や C^{offset} はそれぞれ発音時刻や消音時刻がどの程度ずれやすいかを表す定数である．

3. 解析アルゴリズム

3.1 2次元 LR パーサ

PCFG を解析する手法には一般化 LR 法や CYK 法等がある．しかし，これらの手法は時間方向の順序関係が重要であるため，音高方向に拡張した 2 次元 PCFG には単純に適用することが出来ない．

例えば，図 3 に示すようにある 2 拍の時間のうち最初の拍に 4 分音符が二つ (a と b とする)，次の拍に 4 分音符が二つ (c と d とする) あるという単純な場合について考える．この場合においても，

- 1) a と c, b と d がそれぞれ時間方向の構造を持つ場合
 - 2) a と d, b と c がそれぞれ時間方向の構造を持つ場合
 - 3) a と b, c と d がそれぞれ音高方向の構造を持つ場合
- と 3 種類の場合を持ってしまう．これが，音符数が多く

なっていくときに、どの音符同士が時間方向の文法適用により分割されたかの並びをすべて数え上げようとする指数関数的に場合の数が増大する。そこで、様々な並びを考慮に入れつつもビームサーチで確率が低い候補を切り落とすことで、計算量を削減するという方針を考える。このため、時間的に順番に解析していき、解析途中で確率を評価することが出来る PCFG の解析手法である一般化 LR 法 [6] を応用することを考える。

一般化 LR 法は、現在入ってきた入力とこれまでの構文解析の履歴からどのような解析を行うかという規則 (LR テーブルという) をあらかじめ計算しておき、入力を逐次的に読み込みながらその LR テーブルに従い解析を行うので、非常に効率のよい構文解析手法である。一般化 LR 法の解析器 (以下パーサと呼ぶ) は構文解析の履歴に対応する状態をスタックとして持ち、その先頭の状態と入力から次の 4 つの行動を選択する。

- 1) shift: 現在入ってきた入力を取り込み、次の状態をスタックする。
- 2) reduce: 文法規則の適用を行い、対応する状態をスタックから削除し、スタックの先頭の状態と対応する文法規則に従い次の状態をスタックする。
- 3) reject: 文法解析が行えないので、当該仮説を消滅させる。
- 4) accept: 文法解析が成功する。

また、どの行動を行うかが複数存在する場合はその分だけ仮説を増やしたのち、それぞれ計算する。このため、一般には仮説が指数関数的に増大し、計算が困難になるが、解析途中の確率値などを規準にビームサーチを行うなどの手法により、近似的に計算が可能になる。

この一般化 LR 法を 2 次元 PCFG に応用するために、本研究ではスタックを複数内包するマルチスタックを持つ 2 次元 LR パーサを提案する。まず、LR テーブルの 2 次元 PCFG への拡張として、“synchronization reduce” という動作を新たに導入する。例として、次のような音高方向の文法規則を考える。

$$A \rightarrow B C. \quad (4)$$

このとき、この音高方向の文法規則を次の時間方向の文法規則として取り扱う。

$$A \rightarrow B, \quad (5)$$

$$A \rightarrow C. \quad (6)$$

ここで、式 (5)、(6) に対応する “reduce” の動作に対して “synchronization reduce” を割り当てておく。そして、例えばマルチスタック内のあるスタックが式 (5) に対応する “synchronization reduce” が選択されると、他のスタックが式 (6) に対応する “synchronization reduce” が選択されないか調べて、もしそのようなスタックがあるならば、その二つのスタックを統合し、そのようなスタックが存在し

なければ “reduce” の動作を選択する。

次に、入力が入ってきたときに 2 次元 LR パーサとしては次の 2 つの行動を選択する。

- 1) マルチスタック内の各スタックに入力を割り振る。
 - 2) スタックを新規作成し、そのスタックに入力する。
- また、マルチスタック内の各スタックは入力とそれぞれの状態に応じて、一般化 LR 法のスタックと同様の “shift”, “reduce”, “reject”, “accept” に加え、上述の “synchronization reduce” の動作を LR テーブルに従い選択していく。ここで、1) の割り振り方や、1) と 2) の選択の仕方は任意性があり、また、各スタックが行う行動にも任意性がある場合があるので、一般化 LR 法と同様に仮説を増やしたのち、それぞれ計算し、ビームサーチによる計算量の削減を行う。

また、同じ声部の連続する音は音高が近い場合が多いので、1) の入力の割り当ての時に、そのスタックに直前に入ってきた音の音高からどれだけ離れている音が入力されるかを確率で表し、近い音高が連続しやすいようにする。このときの確率を

$$p_{i,j}^{\text{pitch}} \propto \mathcal{N}(i; j, (\sigma^{\text{pitch}})^2) \quad (7)$$

と設計する。ただし、 i は新しく入ってきた音高、 j はスタックに直前に入ってきた音の音高を表す数値で、半音の間隔が 1 となっている。また、 C^{onset} や C^{offset} はそれぞれ発音時刻や消音時刻がどの程度ずれやすいかを表す定数である。

3.2 有限状態空間内を遷移する 2 次元 LR パーサ

3.1 節で示した 2 次元 LR パーサに入力されるシンボルは楽譜上の発音時刻、消音時刻であるべきであるので、2.2 節で示したテンポや発音時刻、消音時刻の生成モデルを考えると、次の解析アルゴリズムが考えられる。2 次元 LR パーサのマルチスタックが Fig. 2 で示される状態空間内を単位時間ごとに遷移していき、MIDI 信号上の発音時刻や消音時刻が観測されると、現在いる状態に対応する楽譜上の発音時刻、消音時刻を推定し、それを 2 次元 LR パーサの入力とし解析する。

つまり、全体の挙動として、以下ようになる。

- 1) 単位時間が進むごとに状態を遷移する。遷移した結果、各状態に存在する仮説の数がビーム幅を超えた場合、各仮説の確率値の大きいほうからビーム幅の分だけ残し、他を削除する (ビームサーチ)。
- 2) 発音時刻が観測されると、現在の状態から楽譜上の発音時刻を推定して、2 次元 LR パーサに入力する。2 次元 LR パーサは入力をマルチスタック内の各スタックに対し、それぞれに入力した場合についてスタックを新規作成した場合について、仮説を複製してそれぞれ計算する。各スタックは入力に従い解析を行い、“reject” が起きればその仮説を削除し、“synchronization

reduce”の選択が起きたら、他のスタックが対応する“synchronization reduce”の選択が起きるかどうかをチェックし、起きるならばスタックの統合を行い、解析を進める。すべての仮説が解析を終えると、現在の確率値が大きいほうからビーム幅の分だけを残し、他を削除する(ビームサーチ)。

- 3) 消音時刻が観測されると現在の状態から楽譜上の消音時刻を推定して、2次元 LR パーサに入力する。2次元 LR パーサは入力をマルチスタック内のスタックに対し、対応する発音時刻を保持しているスタックに入力する。以降の処理は発音時刻の処理と同様である。ただし、消音時刻と発音時刻の順序が入れ替わることがあるため、2) の処理のとき、入力しようとするスタックの最後に入力された音の消音時刻が処理されていないければ、強制的に3) の処理を行い、2) の処理を行う。
- 4) すべての信号を処理した後は各仮説が“accept”するかどうかをチェックして、“accept”したもののうち確率が最大のものを採譜結果とする。

4. 採譜実験

提案手法の動作を確認するために、まず、簡単な例として、F. Beyer 作曲 Op. 101 の No. 51 を人間が演奏した MIDI 信号の解析を行った。次に、C. Debussy 作曲 “Arabesque No. 1” の 6~9 小節を人間が演奏した MIDI 信号の解析を行った。この曲は 2 声部あり、上パートが三連符、下パートが八分音符となっているため、声部を分離せず発音時刻や消音時刻の分布から楽譜を推定する手法には不向きな曲となっており、声部の分離を同時に行う本手法の有効性を確かめるのに適していると考えられる。この曲に対して、比較対象として、武田らによって提案された手法 [4] を用いた。この手法は声部を分離せず、時間方向の 1 次元に射影した手法であるため、この曲においてはうまく推定されないことが予想される。以下この手法を従来法と表記する。

提案手法の各種パラメータは、状態遷移する単位時間を 30 ms、1 拍の長さの最小値を 0.3 s、最大値を 1.5 s、式 (1) の状態遷移確率が比例する正規分布の標準偏差と前の拍の長さの比 C^{trans} を 0.03、式 (2) のどの状態が発音時刻を出力するかを表す確率が比例する正規分布の標準偏差と属する拍の長さの比 C^{onset} を 0.02、式 (3) のどの状態が消音時刻を出力するかを表す確率が比例する正規分布の標準偏差と属する拍の長さの比 C^{onset} を 0.5、式 (7) のピッチに関する標準偏差 σ^{pitch} を 7 半音、拍の最初の状態におけるビーム幅を 5000、それ以外の各状態におけるビーム幅を 3000、スタックが新規作成される確率を 0.0001、声部の上限を 5 とした。さらに、声部が多くなりすぎないように各仮説の声部数 V に対して $0.001^{(V-1)}$ をそれぞれの仮説の確率に乗算して各仮説の確率として評価した。また、2次元 PCFG の各文法適用確率は人手で適当に設定した。

まず、F. Beyer 作曲 Op. 101 の No. 51 の正解となる楽譜と提案手法の採譜結果を Fig. 4 に示す。比較的採譜が簡単と考えられる曲であるため、2、10 小節に少し間違いがある程度で、非常に正解に近い結果を得ることができた。この間違いは、演奏の表現や物理的な制約により、消音時刻が楽譜上の時刻よりも早く生じていることに起因していると考えられる。解決の方針としては、消音時刻の生成確率を楽譜上の時刻よりも早い状態に関しては大きめに設計することや、付点 4 分音符と 8 分音符の組み合わせを 2 分音符とみなすような文法の設計が考えられる。次に、C. Debussy 作曲 “Arabesque No. 1” の 6~9 小節の正解となる楽譜とそれぞれの手法の採譜結果を Fig. 5 に示す。採譜結果を見ると、予想されたとおり従来法ではうまく推定することが出来ず、一方、本手法では 2 声部がうまく分離され、それぞれのリズムも多少の間違いはあるものの、非常に正解に近い結果を得られた。このような間違いは、前後のリズムがどのようなものが現れたかを用いることで推定が可能と考えられるので、文法の精緻化を行うことにより改善が見込める。

5. まとめ

本稿では、リズムとテンポの不確定性を含む MIDI 信号からの自動採譜という問題に対し、リズムやテンポ、発音時刻、消音時刻の生成過程を考え、統合的に解くことによりその解決を試みた。そして、リズムの生成モデルとして 2次元 PCFG を用い、その解析アルゴリズムとして 2次元 LR パーサを提案した。実験により、多声部でリズムが複雑なものな曲に対し、本手法の採譜結果を示した。今後の課題として、現在は人手で与えられている 2次元 PCFG の文法適用確率を学習することや、タイなどの複雑な楽譜への対応、文法の精緻化と多重音解析のモデルと合わせることで音響信号からの自動採譜システムの構築が挙げられる。

謝辞 本研究は JSPS 科研費 26240025 の助成を受けたものです。

参考文献

- [1] Klapuri, A., et al. “Signal processing methods for music transcription,” Springer, 2006.
- [2] Raphael, C. “Automated Rhythm Transcription,” Proc. of ISMIR. 2001.
- [3] Cemgil, A. T., et al. “Monte Carlo methods for tempo tracking and rhythm quantization,” JAIR. Vol. 18, No. 1, pp 45 – 81, 2003.
- [4] 武田ら, “確率モデルによる多声音楽演奏の MIDI 信号のリズム認識,” 情処論, Vol. 45, No. 3, pp.670 – 679, 2004.
- [5] Kameoka, H., et al. “Context-free 2D tree structure model of musical notes for Bayesian modeling of polyphonic spectrograms,” Proc. of ISMIR. 2012.
- [6] Tomita, M. “Efficient parsing for natural language: a fast algorithm for practical systems,” Vol. 8. Kluwer Academic Pub, 1985.

Piano

$\text{♩} = 120$

6

11

16

This block contains the first system of musical notation for a piano piece. It consists of four systems of staves, each with a treble and bass clef. The first system starts with a tempo marking of quarter note = 120. The notation includes various rhythmic patterns such as eighth and sixteenth notes, and rests. Measure numbers 6, 11, and 16 are indicated at the beginning of their respective systems.

(a) 正解楽譜

6

11

16

This block contains the second system of musical notation, which is a transcription of the same piece as in (a). It follows the same structure with four systems of staves and measure numbers 6, 11, and 16. The notation is visually identical to the original score in (a).

(b) 提案手法による採譜結果

図 4 F. Beyer 作曲 Op. 101, 51 番の正解楽譜 (a), 提案手法による採譜結果 (b) .

Piano

(a) 正解楽譜

(b) 提案手法による採譜結果

(c) 従来手法による採譜結果

図 5 C. Debussy 作曲 “Arabesque No. 1” の 6~9 小節の正解楽譜 (a), 提案手法による採譜結果 (b), 従来法による採譜結果 (c) .