隠れマルコフモデルによる音楽リズムの認識

大槻知史[†] 齋藤直樹[††] 中井満[††]
下平博[††] 嵯峨山茂樹[†,††]

本稿では，隠れマルコフモデル（HMM）を用いて，人間が鍵盤入力した演奏情報（標準 MIDI ファイル）の発音時刻の間隔から，意図された音価列を復元推定する手法を提案し，実験によりその効果を実証する．人間が音楽演奏する際の物理的音長は，音価に対応する正規の長さから意識的・無意識的に揺らぐため，楽譜入力や自動採譜などでは，楽譜として意図された各音符の音価を正しく推定するのは容易ではない．そこで，連続音声認識の定式化にならって，音楽的な演奏を学習・認識する原理を HMM の手法を用いてモデル化する．さらに，同様の原理により小節線・拍子推定，テンポ変化推定も可能となることを示す．

Musical Rhythm Recognition Using Hidden Markov Model

Tomoshi Otsuki,[†] Naoki Saitou,[††] Mitsuru Nakai,[††]
Hiroshi Shimodaira[††] and Shigeki Sagayama[†,††]

This paper proposes the use of Hidden Markov Model (HMM) for rhythm recognition from musical performance recorded in the standard MIDI file format. Intentionally or unintentionally, physical durations of musical notes in human performances often fluctuate from nominal lengths of the intended notes. Estimating intended note sequences is, therefore, not trivial for computers. In this paper, we formulate the process of understanding and recognizing musical rhythm patterns using HMM similarly to continuous speech recognition (CSR). It is shown that the same principle enables bar line allocation, beat recognition, and tempo estimation.

1. まえがき

楽譜の浄書，MIDI 自動演奏などを目的として，楽譜データをコンピュータに入力することが必要な場面は多い．MIDI 楽器で演奏するだけで，意図する楽譜データが入力できれば大変便利である．さらに，演奏された MIDI データから楽譜の書き起こしが自動化できれば，さらに便利である．

しかし，この問題は単純ではない．たとえば MIDI 鍵盤入力の場合，各音符に関してその音高は正確に得られるが，物理的な音長は MIDI の時間分解能を単位としてほぼ連続的な値として観測され，それを単純に処理しただけでは，意図された音価は得られない．その理由は，意図した音符の音価に対応する正規の音長に対し，ユーザが演奏した音長には長短のずれが含まれるからである．さらに，音響信号からの自動採譜においても，各音符の音高の推定のみならず，これと同様の問題が含まれる．

楽譜入力を扱う市販ソフトウェアでは，この変動を減らすために，メトロノームを用いて演奏テンポを一定にしたうえで音符長をクオンタイズ（quantize，量子化）する機能を持つことが多いが，よほどの熟達者ですら，全音符から 16 分音符までを機械的に正確な整数比で弾き分けるのは困難である．まして，音楽初心者が演奏する場合，テンポや正規の音価に対し忠実に演奏することができない場合が多い．さらに,（楽譜入力を目的としない）音楽的な演奏では，曲のスタイル・表情づけ，演奏者の音楽的意図などにより，テンポや音長は意識的な変動を受ける．**図 1** は，ある市販ソフトウェアによる MIDI データから誤った音価列の推定の例である．演奏者の意図は同図左の楽譜であり，同図右の楽譜は演奏に物理的に忠実ではあるものの，目的に合わない．そのため，実際の楽譜入力はグラフィカルな操作により行われることが多いのが現状

† 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
†† 北陸先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Japan Advanced Institute of Science and Technology

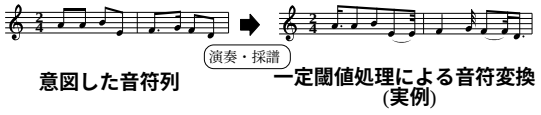

図 1 閾値処理による誤変換の例
Fig. 1 Incorrect results by quantization of note length sequence.

である．

このような，意図した音価に対応する音長からの揺らぎに対して補正する研究はいくつか報告されている．閾値処理をベースとして，ヒストグラム処理による基準拍を設定し，さらにフレーズの終わりは長めになるという音楽的なヒューリスティックルールを付加し，強制を行う手法[2]や，またはテンポ情報を事前に与えて閾値設定に用い，クロック音から音の持続時間の強制を行う手法[3]がある．さらに，各音符長が比例関係にあることに着目した制約として，閾値処理に加えリズムを構文木と捉え文法的な強制を行う手法[1]や，隣接する音長の比が有理数になれば安定するエネルギー関数により，安定するまで処理を繰り返す手法[4]が報告されている．また自動演奏という観点から，演奏情報と楽譜情報との比較から演奏の表情規則を抽出し，その規則により表情づけされた演奏からの採譜システムとする手法[5]などがある．音響信号入力からの自動採譜[8]でも，この問題は扱われており，音響信号から周波数解析・音楽的分析を行い，様々な音楽解釈から楽譜を推定する手法がある[8]．曲のビートを解析するビートトラッキングをマルチエージェントによりモデルベースで音楽的解析を行う報告もされている[6),7]．

一方，訓練を受けた人間ならば，多少の揺らぎがあっても，簡単な音楽ならもっともらしく楽譜化できる場合が多い．これは，人間は常識的なリズムを知識として持っており，それを top-down 的に活用しているからであろう．そのような観点から，本稿では，同種の構造の問題を扱っている連続音声認識分野の方法論の活用を試み，その第 1 段階として，単旋律の MIDI 情報を入力として，その中の発音時刻と音長の情報のみから意図された音価列（直感的には，音楽リズムと理解してよい）を推定する問題を扱う．さらに，演奏テンポ推定，拍子・拍節推定について，その定式化と実験結果について述べる．

2. 連続音声認識問題との同型性

連続音声認識は，近年著しく発達した技術分野の 1 つであり，そのアプローチは音声認識以外の多くの分野でも利用されつつある．基本的な構成要素は，音声

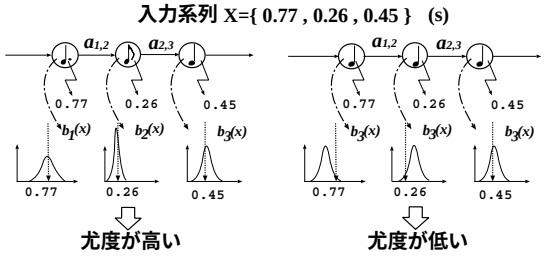

図 2 逆問題としての音価列推定
Fig. 2 Estimation of time value sequence as an inverse problem.

表 1 音声認識とリズム認識の対応
Table 1 Analogy between speech recognition and rhythm recognition.

	連続音声認識	音価系列認識
入力単位	文音声	楽曲
語彙	単語	リズムパターン
単位モデル	音素	音符
隠れ状態	音響イベント	
観測値	スペクトル列	物理的音長列

分析，音響モデル，言語モデル，探索過程の 4 つである．音声分析は，入力音声から有効な特徴ベクトル時系列へ変換する．音響モデルは，その部分時系列に対して音素仮説ごとに尤度を計算することができるよう確率モデルの一種である隠れマルコフモデル（Hidden Markov Model，HMM）[9),10]により音素をモデル化する．言語モデルは，文法や音素列の確率モデルなどにより，許される発声内容を規定する．探索過程は，言語モデルの拘束下で許されるあらゆる音素系列の仮説の中で，尤度が最大となるものを効率良く求める．このようなトップダウン的な考え方を，本稿でも利用する．

本稿では，楽譜上の音符の（整数関係にある）正規の長さを「音価」（time value；時価ともいう）と呼び，それが演奏されて音の物理的長さとして観測されたものを「音長」と呼ぶことにする．これは，音声認識における音素と特徴量の関係に類似している．演奏は，意図された音価系列が揺らぎを持つ音長系列に変換される過程であるとみなす．本問題はその逆問題として音長系列から音価系列を推定する（**図 2**）問題と考えられ，連続音声認識とは**表 1** のように同種の問題である．音声認識における音素を音符の音価に対応づけ，語彙や文法制約を音価列の制約に対応づければ，問題を解くアルゴリズムも対応づけができる．

具体的には，言語モデルに相当する音価系列モデルは状態遷移ネットワークで表現し，楽曲データにより

学習を行う．音素モデル（音響モデル）に相当する音長の変動モデルは隠れマルコフモデル（Hidden Markov Model，HMM）[9),10)]により表現し，実際の演奏データを用いて学習する．音声認識での解探索は，両者を合わせて展開した巨大な HMM の中で Viterbi 探索によって行われるのと同様に，音価系列の推定は，演奏された音長系列がモデルから生成されるあらゆる音価の遷移系列の中で，最も尤度が高い遷移系列を Viterbi 経路探索によって求めることによって行う．音素は自己ループを持つ隠れ状態を複数個用いて表現されることが多いが，音価の場合はそれを自己ループを持たない隠れ状態 1 個のみで表現でき，考え方もアルゴリズムも簡単になる．

3. 音長系列生成過程の確率モデル化

3.1 音価系列モデル

音長に揺らぎがある演奏でも，聴き手には意図した音価の列（さらに，時には伸縮の意図も）が伝わるのはなぜか．その理由の 1 つは，聴き手が出現しうる音価列に関する常識を持っているからであろう．たとえば図 1 右のような楽譜は理論上は可能ではあるが常識に合わない．そこで，聴き手や音楽家の常識をモデル化するために，本手法では音楽的な制約として音価の系列をモデル化する．これは音声認識における言語モデルあるいは文法に相当する部分である．

ここでは，言語モデルがしばしば状態遷移ネットワーク（有限状態オートマトン）で表現され，これを展開すると音素のネットワークとして理解できることにならい，音価の系列の生成源を確率的状態遷移ネットワークで表現する．**図 3** に示すように，各状態はある音価を持ち，ネットワーク全体は許される音価列の全体を表現するものである．各状態には排他的に任意に番号づけがなされ，i 番の状態から j 番の状態への遷移確率は a_{ij} で表現される．音価列の t 番目の音価を生成する状態番号を q_t とすると，ネットワーク上のある状態遷移経路 $Q = (q_1, q_2, \cdots, q_T)$ は，あるリズムパターンを表現し，その生成確率 $P(Q)$ は対応する状態遷移確率の積として与えられる．本稿では，ネットワーク構造として以下の 2 種類のタイプを扱い，そのどちらか一方を用いる（両者の同時使用はしない）．

3.1.1 リズム単語モデル

連続単語音声認識に対応づけられるモデルである．**図 4** に示すように，出現する可能性のある短い音価系列，すなわち「リズム単語」を定義し，単語に相当するリズムパターンの連鎖により曲が成立していると仮定するモデルである．このモデルでは，単語内で隣接する状態 i から状態 j への遷移確率 a_{ij} は，すべて 1 である．また，あるリズム単語の最後の状態 i からあるリズム単語の最初の状態 j へ遷移する確率 a_{ij} は，これらのリズム単語が接続する確率を表している．同様の考え方で，ある状態 i が最初の状態として選ばれる確率 π_i も定義される．

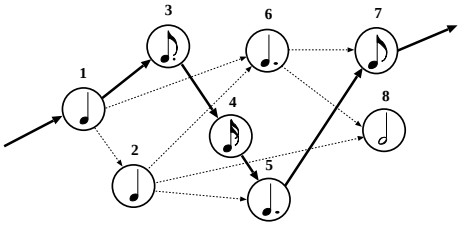

図 3 可能な音価列を表現する状態遷移ネットワーク
Fig. 3 State transition network representing possible time value sequences.

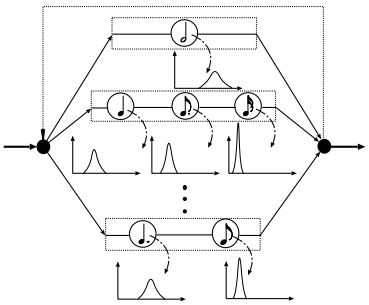

図 4 2 拍単位のリズム単語モデル例
Fig. 4 An example of 2-beat pattern rhythm model.

これは，モデルに含まれているリズム単語の連鎖のみ認識できる点で，モデルとしての拘束力は強いが，未知のリズム単語は扱えない．この点は，音声認識における未知語の問題と同様である．

3.1.2 *n*-gram モデル

音素タイプライタ方式の音声認識に対応づけられるモデルである．Bigram(2-gram) の場合は，**図 5** に示すように，任意の音価 i に任意の音価 j がそれぞれ確率 a_{ij} で後続する．この場合の状態数は，対象とする音価の種類数に一致する．任意のリズムパターンに対処でき，原理的に未知パターンが存在しない利点があるが，音価列を規定する文法としての拘束力は弱い．そこで，$n = 3$，4 とし，trigram(3-gram)，quadgram(4-gram) 遷移確率を用いて，それぞれ 2，3 状態の過去の履歴も考慮することにより，拘束力を強めることができる．これは，状態遷移ネットワークにおいて，単純マルコフ遷移でなく多重（すなわち $(n-1)$ 重）マルコフ遷移を考慮することに相当する．

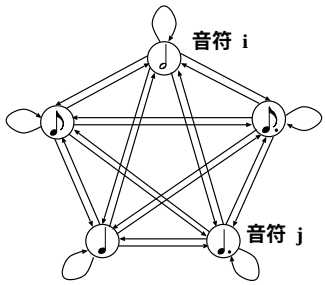

図 5 bigram のモデル例
Fig. 5 An example of bigram model.

表 2 音価列パターンの出現頻度例（4/4 拍子）
Table 2 Examples of frequent time value sequences.

頻度順	1 小節単位	[%]	2 拍単位	[%]
1 位		4.9		16.7
2 位		4.7		12.4
3 位		4.2		11.9
⋮	⋮	⋮	⋮	⋮

3.1.3 モデルパラメータの学習

以上のモデルのモデルパラメータは，楽曲データから学習することができる．これは，人間の音楽経験に基づく常識の形成にたとえられる．

実際に，まず童謡・民謡・歌曲[11)~13)]を対象に 4/4 拍子の曲 88 曲よりリズム単語の統計を得た．リズム単語の単位として 1 小節単位と 2 拍単位の 2 種類を作成し，リズム単語の種類は 1 小節単位 267 種類，2 拍単位 137 種類が得られた．また 3/4 拍子についても同様に 25 曲から統計をとり，1 小節単位 68 種類が得られた．**表 2** に例を示す．これらを用いて，あるリズム単語からフレーズが開始する確率，あるいはそれがアウフタクト単語である確率，リズム単語間の連鎖確率などを学習することができる．これらから，一般的に，あるリズム単語の最後の状態 i からあるリズム単語の最初の状態 j へ遷移する確率 a_{ij} を求めることができる．

次に n-gram$(n = 2, 3, 4)$ 遷移確率を得るため，携帯着信メロディ用の単音のクラシックデータを用いて，全 130 曲，50,000 音程度の学習用の遷移頻度の統計を得た．n-gram 確率モデル場合は，各状態が音価と対応するので，この統計データから，たとえば音価 i からフレーズが開始する初期確率 π_i や音価 i から音価 j への遷移確率 a_{ij}，などの n-gram 遷移確率の値を得た．ただし，少ない学習データに由来する推定誤差を軽減するため，n-gram 確率には 1-gram(unigram) から $(n-1)$-gram までの確率値との線形和による補正（スムージング）を施した．

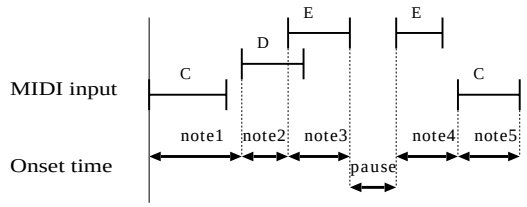

図 6 IOI 処理による音長系列 X の導出
Fig. 6 IOI processing for deriving note length sequence X.

3.2 音長モデル

3.2.1 音長の定義

まず,（演奏情報から得られる）音長の定義をしておく．**図 6** に示すように，個々の音符が演奏される継続時間は，レガートやスタッカートなどのアーティキュレーションによって変動し，後続する音との間で，音が重なり合いあるいは空隙が生じることが多く，音符の音価に対応する音長の物理的観測量としては不適切である．演奏者の音価の反映，あるいは聴取者の音価認知の観点から，さまざまな議論が可能であろうが，本稿では便宜的に発音時刻の間隔（IOI；inter-onset interval）を音長として扱った．

また，この音長が音の継続時間より 0.5（sec）以上長い場合は，その間の無音区間を休符が存在するとみなした．

以下の実験では MIDI キーボードから入力した標準 MIDI ファイルから，以上の処理（以下「IOI 処理」）により，音長あるいは休符長 x_t の系列 $X = \{x_1, x_2, \ldots, x_T\}$ を抽出した．

3.2.2 音長の変動モデル

演奏者は，楽譜として意図した内部状態の系列 Q に相当する音価列から，その演奏者の音楽的な表現（アゴーギグ），演奏の癖，演奏のスキル不足などの原因で，同一の音価の音符でもその物理的音長が変動する．単純化して考えるため，これらを確率変動と見なそう．

状態 k が持つ音価が音長 x で演奏される確率密度を $b_k(x)$ と書く．そのパラメータは，演奏データから学習することができる．これは，人間の音楽経験に基づく音長の揺らぎの常識の形成にたとえられる．ネットワーク上の経路 Q は音価の列（リズムパターン）に対応するので，Q が与えられた場合に音長系列 X が観測される確率を $P(X|Q)$ と書く．

図 7 に，テンポを 96 に保ったのべ 50 人の演奏のデータから得られた，4 分音符，8 分音符，符点 4 分音符の音長ヒストグラムの例を示す．横軸（tick）は各音符の分解能を表す．今回は 4 分音符の音価を 480

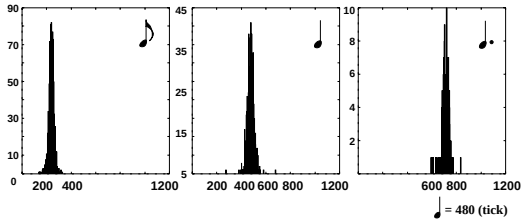

図 7 テンポ指定時の演奏の音長分布（1/960 秒単位）
Fig. 7 Distribution of note lengths with the tempo specified.

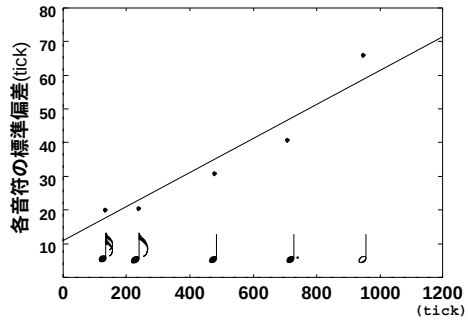

図 8 各音符の音長の平均と標準偏差の関係
Fig. 8 Relation between mean and standard deviation values of note lengths.

（ticks）として統計を得た．なお，本稿で対象とする音価の種類は 16 分音符を分解能とする 16 種類（最長全音符）に 3 連 16 分音符，3 連 8 分音符，3 連 4 分音符を加えた計 19 種類であり,「休符挿入」の場合は，上記の長さの休符も考慮した．

本稿では，各音符の音長の分布を正規分布で近似する．さらに限られた量のデータから分布パラメータを得るために，正規分布の平均 μ は各音符の正規の長さ（音価に対応）とし，標準偏差 σ は音価に比例する分と固定分の和 $\sigma = \alpha\mu + \beta$ の形で与えられると仮定した．α は，統計結果から，各音符の音長の演奏の際の偏差が音価が長いほど広がるという実験事実に基づいた音符間での標準偏差の相違を示し，β は人間の演奏内に含まれる固定分の物理的なずれを表す（**図 8**）．

図 7 から最小二乗法で得られた実験式は $\sigma = 0.05\mu + 0.011$（秒単位）であった．これを**図 9** に示す．

3.3 音長系列生成確率

上記の 2 階層の確率モデルにより，内部状態系列 Q に相当する音価列を意図し，その演奏が音長時系列 X として観測される確率が得られる．すなわち，音長系列 X の生成確率 $P(X|Q)P(Q)$ は上記の 2 つの確率の積で表すことができ，

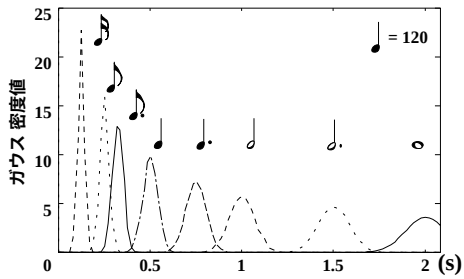

図 9 各音符の音長変動モデル
Fig. 9 A model of fluctuating music note length.

$$P(X|Q)P(Q) = \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(x_t) \qquad (1)$$

となる．q_t は，t 番目の音符を生成する状態番号であり，すでに述べたような状態遷移ネットワークで許される状態遷移経路を表現している．ただし，$n > 2$ の場合の n-gram モデルの場合は，状態遷移確率 $a_{q_{t-1}q_t}$ は n 状態間の多重遷移確率に置き換えるものとする．

4. HMM を用いた音価列推定

本章では，前章で論じた音価系列から音長系列が生成される確率モデルの逆問題として，音価系列を推定する問題を考える．

4.1 逆問題としての音価列推定

演奏された音長系列 X がある内部状態系列（音価列）Q を意図した結果である確率（事後確率）$P(Q|X)$ は，Bayes の定理によって

$$P(Q|X) = \frac{P(X|Q)P(Q)}{P(X)} \qquad (2)$$

と表される．$P(X)$ は経路 Q に依存しないので，異なる経路仮説ごとの $P(X|Q)P(Q)$ を比較して，最大値を与える経路を求めることによって，最ももっともらしい状態系列（音価列）Q^* を推定することができる．式 (1) により，そのような経路（Viterbi 経路）は，状態遷移ネットワークで許されるあらゆる経路 $Q = (q_0, q_1, \ldots, q_T)$ について

$$Q^* = \underset{\{q_0,q_1,\ldots,q_T\}}{\operatorname{argmax}} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(x_t) \qquad (3)$$

を求めることによって得られる．**図 10** に概念的に示すように，よりもっともらしい仮説に対しては (3) 式右辺のこの音長系列 X を生成する確率を表す値は大きくなる．

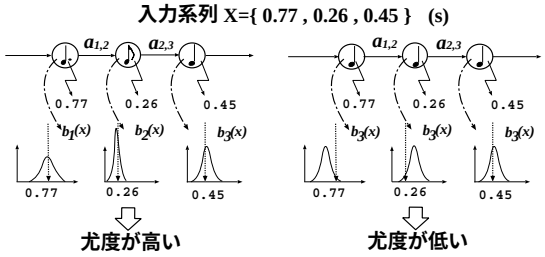

図 10 HMM による尤度計算の例
Fig. 10 An example of likelihood calculation with HMM.

4.2 リズム単語モデルの場合の Viterbi アルゴリズム

リズム単語を単位とするモデルの場合は，連続単語認識と同様にして解を得ることができる．リズム単語内では経路の分岐がなく，リズム単語間で連鎖の確率が与えられる．全体では大きな HMM と考えることで，Viterbi アルゴリズムが適用できる．

4.3 n-gram モデルの場合の Viterbi アルゴリズム

音価列 bigram(2-gram) モデルを用いる場合は，ergodic HMM における Viterbi 経路探索により解が得られる．

一般の n-gram($n = 3, 4 \ldots$) の場合，bigram の場合と同様に，n-gram 遷移確率 $a_{ijk\ldots}$ を用いて生成確率の定式化を行い，その式を用いて表される尤度を最大にする音価列を得るのだが，この問題は bigram の Viterbi 計算に帰着できる．

たとえば 3-gram の場合，bigram の場合の状態空間 S の直積空間 $S \times S$ を状態空間とする HMM を考えることにより，この方法の計算量は bigram の場合のたかだか定義した状態数（倍）である．次節の実験では $n = 3$，4 の場合に，この Viterbi 計算を用いた．

しかし，n が大きい場合はこの方法では計算量やメモリ量の点で限界がある．そこで $n = 3$，4 の場合に，上で述べた n-gram の Viterbi 計算を直接適用する手法だけでなく，まず bigram の尤度 N 位までの候補に解を絞り，次に n-gram($n = 3, 4$) 遷移確率を用いて再ソートを行う 2-pass 手法も用いた．この方法では，N-best アルゴリズムと呼ばれる効率良く上位 N 個の最適解を求めるアルゴリズムが利用でき，N を選ぶことにより精度を落さずに計算量やメモリ量を節減できる．

4.4 音価列推定実験

本節では，実際の演奏データから演奏者の意図した音価列を復元推定する実験を行い，市販ソフトや単純な閾値処理の結果と比較することで，本章で導入する HMM を用いたモデルの有効性を検証する．

4.4.1 実験のデータの流れ

MIDI データの入力には，MIDI キーボード（YAMAHA CBX-K2）を用い MIDI 音源（YAMAHA MU2000 TONEGENERATOR）を通して PC に入力する．演奏収録ソフトとしては，YAMAHA XGworks ver. 4.0 を用いた．次に，得られた MIDI データの発音時刻の間隔から音長系列 X を導出した．この X に対し，学習で得た音価系列モデルおよび音長変動モデルを用いて Viterbi 探索を行い，音価系列 Q^* を出力した．

4.4.2 実験対象曲および演奏条件

実験対象曲としては，リズム単語モデルでは，その効果が分かりやすいように，比較的単純なリズムパターンが多い童謡など 16 曲の旋律を用いた．一方，n-gram モデルでは，三連符なども含む多少複雑なリズムを扱ってその効果を確かめるために，クラシック曲 8 曲の冒頭部とした．なお対象曲は，学習に用いた曲に含まれていない．被験者は，何度も弾き直さないと正しく演奏できない者から音楽大学卒業生まで，幅広い演奏スキルを持つ 19 人である．

演奏データとしては，認識率の計算のために演奏誤りがない（演奏した音の数と楽譜上の音符の数が一致している）ものを用い，実験には

演奏条件 1： メトロノームを用いた，テンポにできるだけ忠実な演奏

での入力を用いた．ただし，演奏テンポは既知とし，それに基づいたモデルを用いる．

4.4.3 評 価 方 法

本章では，演奏された各音長が正しい音価に変換されているかを評価するために，以下の式で認識率を与えた．

$$\text{accuracy} = \frac{N - \text{sub} - \text{del} - \text{ins}}{N} \times 100 \quad [\%] \quad (4)$$

ただし，N は総音符数，sub は音価の置換誤り数，また del，ins は「休符挿入」条件時の休符の脱落，挿入誤り数である．

4.4.4 音価列推定結果

まず，ある一定テンポ [演奏条件 1] の演奏に対し，市販ソフトによる楽譜化と bigram の HMM を用いた場合の楽譜化の比較を行った．**図 11** に示す楽曲の演奏に対し，市販ソフト（YAMAHA XGworks）では **図 12** のような不適切な出力となる一方，HMM を用いた場合は，小節線は入っていないものの音価としては 3 連符を含めて，**図 13** のように末尾の音符の音価以外はすべて正しい楽譜が得られた．

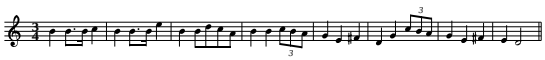

図 11 「Brahms 交響曲 2 番 3 楽章」の冒頭の旋律（装飾音を省略し，単純化してある）
Fig. 11 A testing phrase from the 3rd movement of Brahms' symphony No.2.

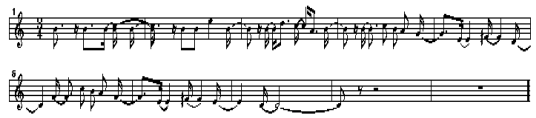

図 12 「Brahms」の演奏の XGworks による楽譜化結果
Fig. 12 The score obtained by XGworks from performance of 'Brahms'.

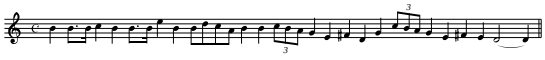

図 13 「Brahms」の演奏の HMM による楽譜化結果
Fig. 13 The score obtained by HMM from performance of 'Brahms'.

表 3 bigram HMM とリズム単語 HMM の認識率の比較 [単位：%]
Table 3 The recognition rate of bigram HMM and rhythm-vocabulary HMM [%].

method	休符挿入	休符無視
閾値処理 (XGworks)	40.70	85.86
リズム単語 HMM	59.65	97.26
bigram HMM	53.73	87.39

次に，童謡などの 16 曲の一定テンポの演奏 [演奏条件 1] に対し，リズム単語モデルと bigram モデルの HMM のそれぞれを用いて実験を行った．その結果，**表 3** に示すように，リズム単語モデルの場合の認識率は，bigram モデルの場合を上回った．これは，未知リズム単語が出現しないかぎり，リズム単語モデルの方が文法的拘束力が強く働くためであると理解できる．

連打音，スタッカート，フレーズ境界などで自然に生じる音間の空隙は，休符と完全に区別することは困難である．そこで，出力された休符をすべて先行音の延長として置き換えた性能評価（休符無視）も行った．

また，n-gram を用いたクラシック曲の一定テンポの演奏（[演奏条件 1]）に対して音価列推定を行った．市販ソフトの楽譜化における音価認識率は 40%程度と低いため比較対象として適切でない．本手法の効果を評価するために，IOI 処理後に閾値処理する方法の認識率を性能の基準として用いた．

表 4 に IOI 処理後に閾値処理した場合（QUANT），2-，4-gram の HMM を用いた場合（2-，4-HMM）の 3 通りの認識率を比較した結果 [演奏条件 1]（被験者 19 人）を示す．表 4 のように，bigram の HMM の認識率は，1 曲を除いて，閾値処理のみの場合の結果と同等以上であった．また，16 分音符の音価を分解能とする閾値処理では検出できない 3 連符を，HMM の場合は認識した．実際，3 連符を含む表 4 の下 3 曲に対しては，特に高い誤り削減率が見られた．

表 4 音価列推定結果．演奏条件 1 で，IOI 処理後の閾値処理による各曲ごとの正解率（QUANT）と，各手法（2-HMM, 4-HMM）の閾値処理からの誤り削減率．単位：[%]
Table 4 Results of time value estimation. Accuracy by quantization (QUANT) and error reduction rates by bigram+HMM (2-HMM) and 4-gram+HMM (4-HMM).

曲目	QUANT	2-HMM	4-HMM
Ave verum corpus	98.4	0	**50**
別れの曲	97.9	−24	−24
もろびとこぞりて	95.4	15	15
アルルの女	86.3	0	**11**
ボレロ	94.6	22	22
アラヴァマ序曲	81.0	**52**	33
Brahms 交響曲 2 番	65.3	49	**61**
くるみ割り人形	60.0	92	92

さらに 4-gram の HMM の場合，bigram の HMM に比べいくつかの場合について認識率が向上し，この結果から n-gram($n = 3, 4$) の導入は有効であると考えられる．

しかし，連鎖確率の学習サンプル量が多く得られない場合は，かえって誤認識の原因となる可能性があり，実際に，「アラヴァマ序曲」においては 3 連 4 分音符に続く 8 分音符 3 個を，3 連 4 分音符とみなす誤認識のために，認識率が低下した．

5. HMM を用いた変動テンポの推定

音楽的演奏意図や演奏スキルによって，音楽演奏のテンポは無意識あるいは意識的に変動することが多いが，従来の閾値処理（クオンタイズ）では適切な楽譜化ができないことが多かった．また，原理的にも，テンポ変動と複雑な楽譜とを区別することは難しい．しかし，人間は変動テンポに追従して音価列を正しく理解できることが多い．本章では，テンポの異なるリズム単語モデルを並列に持つことで，HMM により解決できることを示す．

5.1 一定テンポ/変動テンポ推定問題

既出のリズム単語モデルは，時間情報として各音長がとりうる値を出力確率に対応させたモデル化であるため，あらかじめ定めたテンポの入力のみ解析可能である．そこで，各リズム単語モデルを複数のテンポごとに作成し，入力に対して各テンポごとに並列に尤度

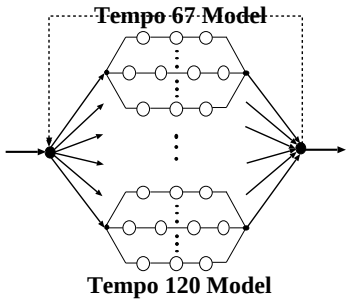

図 14 一定テンポモデル
Fig. 14 Model of constant tempo.

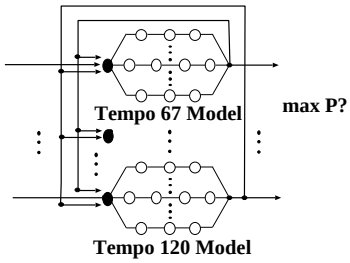

図 15 変動テンポモデル
Fig. 15 Model of fluctuating tempo.

計算を行い，尤度が最大となるテンポを推定結果とすることでテンポによる適用範囲を広げる（**図 14**）．テンポは 67〜120 の間で対数的に 5 分割し，テンポ 67・76・85・95・107・120 の 6 種類を採用した（一定テンポモデル）．

次にテンポの揺らぎが激しい入力に対処するために，**図 15** のように，図 14 の一定テンポモデル間に遷移確率を設け，階層型 HMM を作成する（変動テンポモデル）．このテンポ間遷移確率は，テンポの変わりやすさをモデル化するものであり，今回はヒューリスティックに与えた．これにより，移り変わるテンポに追従した解析が可能になった．

5.2 一定テンポ推定実験

5.2.1 入力データおよび用いるモデル

一定テンポ推定実験の入力は，

演奏条件 2： テンポ指定なしで，できるだけ一定のテンポを保つことを心がける演奏

とし，被験者によく知られていて短く弾きやすく，かつ多様なリズムを含む曲として「もろびとこぞりて」[12] を対象曲として選び，被験者 10 人（10 演奏）を対象とした．用いるモデルは図 14 の一定テンポモデルにより，6 種類のテンポ候補中から演奏されたテンポを 1 つ推定する．また，リズム単語の単位は，2 拍単位とした．

表 5 テンポ推定結果（演奏条件 2．10 曲．A：拍数（38 個）/演奏時間（分），B：一定テンポ HMM による推定）
Table 5 Tempo estimation results (A: beats/min., B: tempo estimated by HMM).

player#	A	B	player#	A	B
1	98.35	95	6	116.41	120
2	93.31	95	7	111.74	107
3	99.20	95	8	99.88	95
4	127.06	120	9	109.25	107
5	106.34	107	10	65.16	67

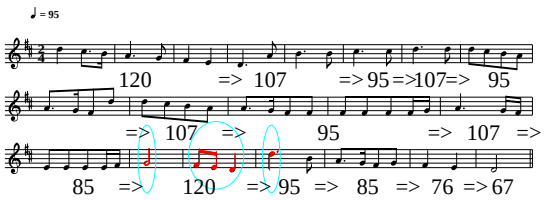

図 16 変動するテンポと音価列の推定（○は誤推定）
Fig. 16 Simultaneous estimation of fluctuating tempo and time value sequence.

5.2.2 評価方法

演奏は奏者の演奏技術による揺らぎ以外の表情づけなどの変動要因は含まないことをふまえ，その曲全体が演奏された平均テンポ（1 分間の 4 分音符の数）を

演奏テンポ = 拍数/演奏時間 (分)

により定義し，比較対象とする．

5.2.3 テンポ推定結果

曲の演奏時間から求めた平均テンポと一定テンポ HMM の選択されたモデル（最も尤度が高いモデル）を**表 5** に示す．6 種のテンポのうち最も近いモデルが選択され，その意味でのテンポ推定率は 100% が得られた．

5.3 テンポ変動認識実験

5.3.1 入力データおよび用いるモデル

前節と同じ入力曲で，

演奏条件 3： メトロノームを用いず，テンポが自由に揺らぐ演奏

に対する実験を行った．モデルは図 15 に示す変動テンポモデルを用いた．最も多く採用されたテンポのモデルを，その曲が演奏された平均のテンポとした．

5.3.2 テンポ変動問題に対する推定結果

図 16 に，意図的に極端なテンポ変動を行った演奏に対するテンポ変動推定実験結果例を示す．尤度最大の 2 拍単位のリズム単語モデル集合間の遷移経路（Viterbi 経路）をたどると，以下のようなテンポモデル間遷移の推定結果が得られた．

Tempo 120（初期モデル）→ 120 → 120 →

107 → 107 → 95 → 107 → 95 → 95 → 107
→ 95 → 95 → 107 → 85 → 120 → 120 →
95 →85 → 76 → 67

極端に遅い演奏個所では，音価は倍にテンポは速めに推定された結果，誤推定が生じた（図 16 の○部分）が，妥当な推定であるとも考えられる．2 拍単位のリズム単語モデルを用いたので小節ごとにテンポが推移するような場合は，小節ごとに誤推定されることがある．

6. HMM を用いた拍節推定

曲を聴いて楽曲の拍子と小節線の位置を推定することは，必ずしも容易ではない．2/4 拍子と 4/4 拍子，3 連符のみの 2/4 拍子と 6/8 拍子など，原理的に演奏からは区別できない場合も多く，さらに意図的に予想をくつがえすような楽譜も可能である．しかし人間は多くの場合，2 拍子系と 3 拍子系との区別や，上げ拍（アウフタクト，弱起）かどうか程度の推定は，比較的正確に行える．本章では，これらの問題も確率モデルの問題と考えて定式化し，解決を図る．

6.1 拍子/開始拍/小節線位置推定問題

演奏から楽譜を復元する場合には，音価列のみならず拍子の推定，開始拍（アウフタクトかどうか）の推定，すなわち小節線をどのように挿入すればよいかという問題を解決する必要がある．これらの問題も，以上に述べた確率モデルによって定式化できる．

拍子特性が顕著に現れるのは，1 小節中に含まれるリズム（音価系列）パターンであると考えられる．そこで 4/4 拍子，3/4 拍子ごとに 1 小節単位のリズム統計を得，各モデルで入力された旋律の尤度を並列計算し音価列を推定する．ここで尤度最大の原理を利用し，尤度が高い遷移系列を求めその系列が 4/4 であるか 3/4 であるかを判定し，拍子推定結果とする．ただし，ここでは，拍子が 1 曲中で変化しないことを前提としている．

小節線推定は，**図 18** のように 1 小節単位のリズム単語モデルを用いる場合は，自動的に行うことができる．また，2 拍単位のリズム単語を用いる場合は 2 拍単位のリズム単語 2 個につき小節線を出力するため，最終状態（リズム単語）から逆算で求める必要がある．

また，リズム単語の中に学習で得た初期確率のみを与えたアウフタクト単語を加えているため，アウフタクトの可能性も含めた小節線位置の推定を行うことができる．

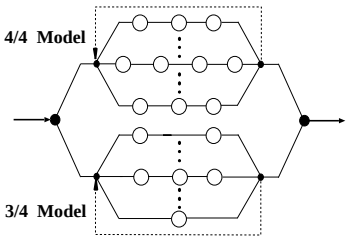

図 17 モデルによる拍子推定
Fig. 17 Model-based measure estimation.

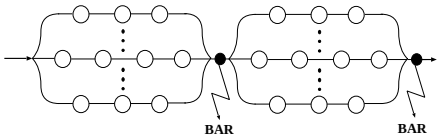

図 18 モデルによる小節線推定
Fig. 18 Model for locating bar lines.

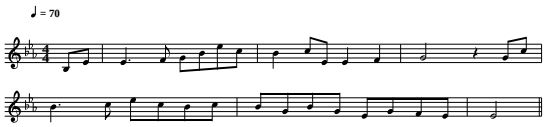

図 19 拍子推定における誤認識例「赤とんぼ」
— リズムパターンの観点からは妥当な解
Fig. 19 An example of beat estimation error in “Akatombo”
— reasonable from rhythm pattern point of view.

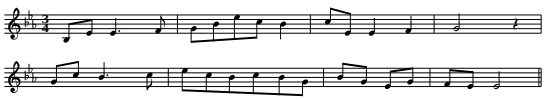

図 20 「赤とんぼ」の正しい楽譜（3/4 拍子）
Fig. 20 The correct score of “Akatombo”.

6.2 拍節推定実験

6.2.1 入力データおよび用いるモデル

拍節推定実験では，**図 17** のモデルを用い，童謡などの 4/4 拍子 10 曲，3/4 拍 10 曲に対し一定テンポ [演奏条件 1] の演奏を入力した．リズムの最小単位としては双方とも 1 小節単位パターンのモデルを用いた．

6.2.2 拍子・小節線推定結果

4/4 拍子については 10 曲すべてについて正しく拍子推定できた．3/4 拍子 10 曲中 8 曲は正しく推定できたが，残る 2 曲は音価列としては正しく推定されたが，拍子は 4/4 拍子と誤推定された．**図 19** に誤推定例を，**図 20** に正しい楽譜を示す．リズムパターンとしては，1 フレーズが 3 小節になっているところに違和感があるが，4/4 拍子と考えても矛盾はない．このような場合の拍子推定は，旋律あるいは想定される和声まで含めたさらに高度な総合モデルが必要となる．

また，拍子が正しく推定された演奏に対しては，ア

ウフタクトも正しく推定できた．今回は，拍子が不変の曲を対象としているため，音価列の推定が正しい結果については，小節線も正しく推定できた．

しかし，拍子を誤推定した場合，小節線は本来の楽譜とは異なる位置に挿入された．また，拍子推定が正しくとも，音価列（リズム単語）の推定誤りによって，正しい位置に小節線が挿入されない場合があった．

7. まとめと今後の課題

本稿では，音楽演奏の音長系列データに対し，連続音声認識の方法論を適用して統合的な確率モデルと最尤経路探索により，意図された音価列推定，テンポ推定，拍子推定，小節線位置推定などが統一的に行えることを示した．このような確率モデルによるアプローチは，従来よく行われたボトムアップの処理，あるいはルールベースの処理に比べて，モデルの学習が可能であり，今後の高い可能性を持つ．本稿の目的はこのような新しい手法の提案にあるため，今回は限られたデータ量による実験結果であるが，今後，学習用に十分な音楽データを整備することにより，さらに高い性能が期待できる．

今後は，ジャンルやスタイルを考慮した（に依存した）リズム単語のモデル学習方法，楽曲フレーズのようなより大きな曲構造を反映したモデル，未知リズム単語への対処（音声認識における未知語対策に対応），リズム単語に依存した音長伸縮特性を考慮した推定（同じく文脈依存モデルに対応），ユーザのスキルや癖を学習するユーザ適応技術（同じく話者適応に対応）などの発展により本法の適用可能性を広げたい．さらに，音響信号入力に対して適用し，自動採譜の一要素技術として用いたい．

参 考 文 献

1) Ronguet-Higgins, H.C.: *Mental Processes*, The MIT Press (1987).
2) 片寄，井口：知的採譜システム，人工知能学会誌，Vol.5, No.1, pp.59–66 (1990).
3) 海野，中西：音楽情景分析における楽音認識と自動採譜，インタラクション 99 予稿集 (1999).
4) Desain, P. and Honing, H.: Quantization of Musical Time; A Connectionist Approach, *Computer Music Journal*, Vol.13, pp.56–66 (1989).
5) 野池，乾，野瀬，小谷：演奏情報と楽譜情報の対からの演奏表情規則の獲得とその応用，情報処理学会研究報告（音楽情報研究会），97-MUS-26-16, pp.109–114 (1998).
6) 後藤，村岡：音楽音響信号を対象としたビートトラッキングシステム—小節線の検出と打楽器音の有無に応じた音楽的知識の選択，情報処理学会研究報告（音楽情報研究会），97-MUS-21-8, pp.45–52 (1997).
7) Goto, M. and Muraoka, Y.: Real-time Rhythm Tracking for Drumless Audio Signals: Chord Change Detection for Musical Decisions, *Speech Communication*, Vol.27, Nos.3-4, pp.311–335 (Apr. 1999).
8) 長嶋，橋本，平賀，平田：コンピュータと音楽の世界，bit 別冊，共立出版 (1998).
9) 中川：確率モデルによる音声認識，電子情報通信学会 (1988).
10) Rabiner, L. and Juang, B.-H.: *Fundamentals of Speech Recognition*, Prentice-Hall (1993).
11) 中学生の音楽 1，2，3，教育芸術社 (1983–85).
12) 楽しく歌おう，神奈川県中学校音楽教育研究会 (1983).
13) 世界名歌 110 曲集 (1)，全音楽譜出版社.

(平成 13 年 6 月 18 日受付)
(平成 13 年 12 月 18 日採録)

大槻 知史

2001 年東京大学工学部計数工学科卒業．現在，同大学大学院新領域創成科学研究科複雑理工学専攻に在籍．

齋藤 直樹

1998 年創価大学工学部情報システム学科卒業．2000 年北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．現在，（株）PFU 勤務．

中井 満（正会員）

1991 年東北大学工学部情報工学科卒業．1993 年同大大学院博士前期課程（情報工学）修了．1996 年同大大学院博士後期課程（電気・通信工学）修了．1996 年北陸先端科学技術大学院大学情報科学研究科助手，現在に至る．博士（工学）．音声認識，文字認識に関する研究に従事．電子情報通信学会，日本音響学会，各会員．

下平 博（正会員）

1982 年東北大学工学部電気工学科卒業．1984 年同大大学院博士前期課程（情報工学）修了．1988 年同博士後期課程修了．同年東北大学工学部情報工学科助手．1992 年北陸先端科学技術大学院大学情報科学研究科助教授，現在に至る．工博．音声，文字，画像の認識処理およびヒューマンインタフェースに関する研究に従事．日本音響学会，電子情報通信学会，IEEE 各会員．

嵯峨山茂樹（正会員）

1972 年東京大学工学部計数工学科卒業．1974 年同大大学院工学系研究科計数工学専攻修士課程修了．同年，日本電信電話公社に入社，武蔵野電気通信研究所にて音声情報処理の研究に従事．1990 年 ATR 自動翻訳電話研究所音声情報処理研究室長として自動翻訳電話プロジェクトを遂行．1993 年 NTT ヒューマンインタフェース研究所にて音声認識・合成・対話の研究開発に従事．1998 年北陸先端科学技術大学院大学情報科学研究科教授．2001 年東京大学大学院工学系研究科のち情報理工学系研究科教授．博士（工学）．1990 年発明協会発明賞，1994 年日本音響学会技術開発賞，1995 年情報処理学会山下記念研究賞，1996 年科学技術庁長官賞（研究功績者表彰）および電子情報通信学会論文賞等を受賞．日本音響学会，電子情報通信学会，IEEE，ISCA，AVIRG 各会員．