2004－MUS－56　(2)
2004／8／2

部分列エージェントモデルによる楽曲の構造認識

横山 博　　平賀 譲

筑波大学大学院 図書館情報メディア研究科
{hiroshi|hiraga}@slis.tsukuba.ac.jp

あらまし

本稿では、楽曲の進行に伴って階層的な構造を形成する音楽聴取の計算モデルを提案し、初期的な実験結果を報告する。提案するモデルはマルチエージェントモデルに基づいており、記号列で表された楽曲の部分列を表現する部分列エージェント、隣接する二つの部分列エージェントを連結する連結エージェント、連結エージェントの妥当性を評価する評価エージェントの三種類のエージェントから成る。これらのエージェント間の相互作用によって全体として一つの解釈を表す階層構造が作られる。実験では、本モデルの手法を拍節解析に適用し、ゆらぎのある音長列から階層的な拍節構造を得た。また得られた構造について検討し、モデルの問題点とその解決法について論じる。

Recognition of Musical Structure based on Sub-sequence Agent Model

Hiroshi Yokoyama　　Yuzuru Hiraga

Graduate School of Library, Information and Media Studies, University of Tsukuba

Abstract

This paper presents a computational model of music cognition based on a multi-agent architecture. The interaction among the agents generate hierarchical structures corresponding to the interpretation of the music. The model consists of three types of agents — *sub-sequence*, *concatenator*, and *evaluator* agents. A sub-sequence agent is attached to each sub-sequence of the piece, while a concatenator represents concatenation of two adjacent sub-sequences. The musical plausibility of each of these agents is evaluated by the evaluators, and the agents selected from competing candidates collectively present a structural description of the music. In a preliminary experiment, the model was applied to a beat recognition task which generates a metrical hierarchy from the input of note onset sequences. Based on the results of the experiment, we discuss the applicability and extinsibility of the model to music cognition.

1　はじめに

人間が音楽を聞く時、それをただ単に音の羅列として聞いているのでなく、互いに関連づけられた、構造を持ったひとまとまりのものとして捉えている。そのような音楽の持つ構造性を扱った研究として、Simon & Sumner[1]、平賀 [2]、平田ら [3][4] などがあげられる。これらの研究では、楽曲の構造を自然に表現し、高次の音楽的操作を可能とするような知識表現が論じられている。

一方、このような構造の認識は一つの曲を全曲通して聞いてからはじめておこなわれるのではなく、楽曲を聞く過程と不可分に結びついている。例えば人間は歌謡曲や童謡を聞きながら「1 番」や「2 番」

などと感じたり、さらに細かい「最初の部分」や「盛り上がる部分」などのようなまとまりをごく自然に感じたりする。しかしながら、音楽聴取時において人間がどのように楽曲構造を認識しているのかは明らかではない。

本稿では、このような楽曲聴取過程において、階層的な音のグルーピング構造を生成する計算モデルを提案する。

2 関連研究

楽曲の進行に伴ってその解釈をおこなう音楽認知モデルは、拍節認識の分野において多く提案されている。

Longuet-Higgins & Lee [5] は、隣接する同じ長さの拍が上位の拍を形成する、という考えに基づき、ゆらぎのない音長の列を入力すると、拍の開始位置と長さを順次出力するモデルを提案した。このモデルは、明確なルールで拍節認識の過程を説明する。一方、唯一の解釈のみを保持するモデルであり、入力自体にあいまいさがあるような場合であっても、一つの解釈しかなされない。また、解析の初期に誤った解釈をすると、そこから回復できなくなるといった問題がある。

マルチエージェントモデルによって複数の解釈を保持するモデルの例として、Chung [6] によるモデルや、Rosenthal [7] によるモデルがあげられる。これらのモデルでは、異なる解釈の可能性が生じるたびにエージェントが複製され、それぞれの可能性を追跡していく方策がとられている。これにより、入力自体にあいまいさがある場合でも妥当な解釈を捨てることなく処理が続けられる。一方、個々のエージェントのレベルに着目すると、それぞれのエージェントは一つの拍節構造を曲全体に対して適用している。したがって、一度誤った解釈をするとそこからの回復が不可能という点では、Longuet-Higgins & Lee のモデルと本質的に変わりはない。

そこで、曲全体に一つの解釈をあてはめるのではなく、曲の部分ごとに解釈をおこない、それらを統合して一つの解釈を作り上げる方法が考えられる。そのようなモデルの例として、Desain & Honing [8] があげられる。このモデルはコネクショニストモデルに基づいており、与えられたゆらぎのある音長の列を量子化された音長の列に変換する。このモデルでは basic cell と sum cell と呼ばれるセル (基本的な処理単位) が、それぞれ一つの音と隣接する複数の音に対応し、その長さを記憶している。そして隣接するセル間で音長比を整数比に近づける処理を繰り返すことで全体としての処理が進められる。なお、このモデルは楽曲の進行に伴って解釈するような実装にはなっていない。また階層的な出力は得られない。

拍節のみではなく音楽全体に関する認識をリアルタイムでおこなうシステムの例として、Rowe の Cypher[9] があげられる。これは対話的に音楽を作曲・演奏するシステムであり、その処理の一環として、MIDI によって入力される演奏データの認識・解析がおこなわれる。Cypher では曲全体の統一した解釈を得ることよりは、むしろ様々な観点から音を組織づけることが重視されている。階層性を持つグルーピング構造としてはフレーズと呼ばれる単位がある。フレーズは複数の下位のイベント (おおむね一つの音に相当する) から成り、その不連続性を示す値があるしきい値を超えた時に生成される。階層は 2 階層であり、フレーズの上位の階層はない。

3 モデルの概要

本稿で提案するモデルは、Minsky の「心の社会」論 [10] に基づき、マルチエージェントシステムとして構成される。

本モデルの対象とする入力は単音列であり、曲データを記号列 P として表す。個々の記号は、音の何らかの属性 (音高、音長、その複合など) を表現したものである。曲の階層的構造は、P をいくつかの部分列に分割し、隣接する部分列間の連結関係を与えることで得られる。そこで、初期的にはすべての部分列およびそれらの間の連結関係を考え、その中から曲の構造を適切に表すものを取捨選択する過程とし

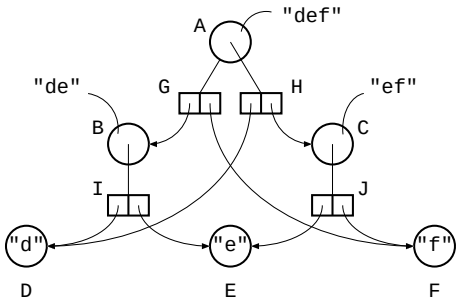

図 1: 部分列エージェント (SA:○) と連結エージェント (CA:□□) の例

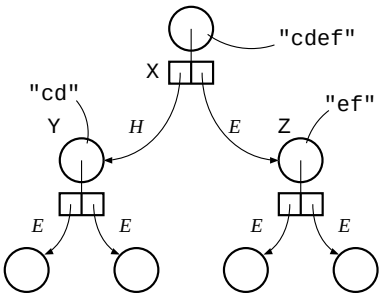

図 2: H–リンクと E–リンクによる “(c d) e f” の表現

て、曲の階層構造認識を捉える。

3.1 モデルを構成するエージェント

各々の部分列、連結関係にそれぞれ一つのエージェントを割り当て、前者を部分列エージェント (SA)、後者を連結エージェント (CA) と呼ぶ。これらのエージェントを取捨選択する評価基準は、一般には複数ある。個々の評価基準ごとに一つのエージェントを割り当て、これを評価エージェント (EA) と呼ぶ。図 1 に、記号列 “d e f” に対する SA と CA の例を示す。A–F の白丸ノードが SA を、G–J の四角ノードが CA をそれぞれ表している。SA に付与されたラベルは対応する部分列を表す。

CA は、ある部分列とそれを分割してできる二つの部分列に対応する三つの SA とのリンクを持つ。そのリンクが CA/SA 間の上位/下位の関係を表す。図 1 では、I から見て B が上位 SA、D と E が下位 SA である。また B から見て G が上位 CA、I が下位 CA である。

長さ k の SA には、部分列を二つに分割する分け方に対応して $k-1$ 個の下位 CA がある。それぞれの SA が下位 CA の一つを「選択」することで全体として階層構造が得られる。図 1 で A–H 間のリンクが A の中心まで達しているのは、A がその下位 CA の G と H のうち、H を選択していることを表している。

記号列の長さを n とすると、SA の総数 (= 部分列の個数) は $n(n+1)/2$、CA の総数は $\sum_{k=1}^{n}\{n-(k-1)\}(k-1) = n(n^2-1)/6$ である。

グルーピング構造を表現するため、CA から下位 SA へのリンクとして、階層型リンク (H–リンク) と列挙型リンク (E–リンク) の二種類を区別する。H–リンクはリンク先の部分列を、入れ子になったひとまとまりの構造単位として扱うことを表す。また E–リンクは部分列を構造単位としては扱わず、そのまま上位の列に埋め込むことを表す。

図 2 にグルーピング構造の表現の例を示す。図中、X–Y 間が H–リンクであるため “c d” は入れ子の列となり、X–Z 間は E–リンクであるため、“e f” はそのまま埋め込まれる。したがって全体は “(c d) e f” という構造を表している。(入れ子になった部分列は括弧をつけて表す。)

CA は下位 SA へのリンクを 2 本持つので、一つの CA が表しうるグルーピング構造の候補は EE、EH、HE、HH の 4 通りある。

3.2 インクリメンタルな構造認識

前節で論じたモデルによる構造認識の処理は、楽曲データ全体をまとめて処理する場合だけでなく、入力データ列を順次処理していくインクリメンタルな処理にも適用できる。これは人間の聴取過程のモデルとして望ましい性質である。

新たな入力があると、対応する SA とそれに接続

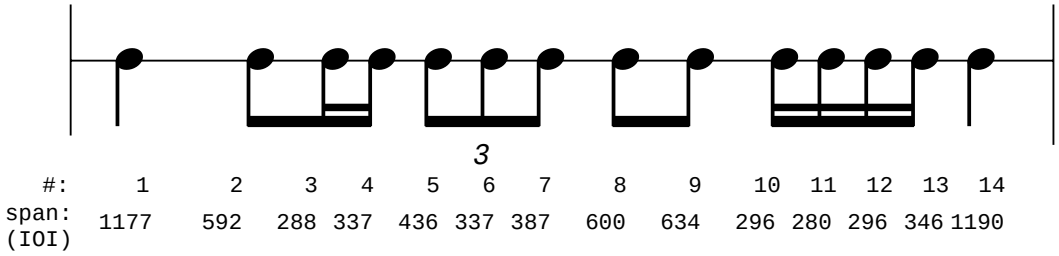

図 3: 実験に用いた音長データ [8]
音長 (span) は上段の楽譜を人間が演奏したものである。

する SA/CA が作られる。入力に対応する SA は全ての上位 CA に対してそれを通知する。例えば図 1 で F が新たな入力に対応する SA であるとすると、それが G と J に通知される。通知を受けた CA は 4 種類のグルーピング構造の候補を作る。EA は 各候補に対し、評価基準にしたがってスコアをつける。スコアが最も高かった候補が CA のグルーピング構造となる。

変更が生じた CA は上位 SA に対してそれを通知する。この通知は図 1 の G–A 間と、J–C 間で起こる。通知を受けた上位 SA は、通知された CA が現在選択している CA よりスコアの高いものであった場合に新たな CA を選択する。

以下同様に処理が続けられ、認識結果はその時点での全体の部分列に対応する SA からリンクをたどることで得られる。

4 適用例:拍節認識

以下に本モデルを拍節認識に適用した例を示す。この例で用いた EA は一つであり、拍節構造のまとまり方を評価する。EA が用いるスコア関数 $sc(X)$ を以下の様に定義する。

$$sc(X) = -\frac{\sigma_x}{\bar{x}} + \sum_{i=1}^{nsub(X)} \frac{span(sub(X,i))}{span(X)} sc(sub(X,i))$$

表 1: 音長列とそのスコアの例

列名	構造	sc
A	1100 900	-0.1
B	600 400	-0.2
$A\ B$	1100 900 600 400	-0.36
$A\ (B)$	1100 900 (600 400)	-0.15

ここで X は、階層化された音長列 (音長リスト) で、$\bar{x}$, σ_x は、X の各要素 (入れ子列の場合はその音長の合計) の平均、標準偏差である。$span(X)$ は列全体の音長を、$nsub(X)$ は X 中の入れ子列の個数を、$sub(X,i)$ は列 X の i 番目の入れ子列を表す。

表 1 に音長列とそのスコアの例を示す。列 A と B は標準偏差は等しいが、A の方が音長の平均が大きくゆらぎの割合が小さいため B と比較して高いスコアとなっている。また A と B を単純に連結した列 $A\ B$ では音長が揃っていないために、スコアは低くなっている。B を入れ子列として連結した列 $A\ (B)$ は、トップレベルでは "1100 900 1000" のように揃った音長となり比較的高いスコアとなっている。このように、関数 $sc(X)$ は音長がよく揃っているほど高いスコアとなる。

図 3 に実験で用いる入力を示す。これは文献 [8] から引用したもので、ゆらぎのある音長列を表す実演奏データである。

入力の各ステップごとに得られた拍節構造の一覧

表 2: 実行結果

ステップ	得られた拍節構造
#1	1
#2	1 2
#3	1 (2 3)
#4	1 (2 (3 4))
#5	1 (2 (3 4) 5)
#6	1 (2 (3 4)) (5 6)
#7	1 (2 (3 4)) (5 6 7)
#8	(1 (2 3)) ((4 5) (6 7) 8)
#9	(1 (2 (3 4))) ((5 6 7) (8 9))
#10	(1 (2 (3 4))) (5 6 7 8 9 10)
#11	(1 2) (3 4 5 6 7) (8 9 (10 11))
#12	1 (2 (3 4)) (5 6 7) (8 9) (10 11 12)
#13	1 (2 (3 4)) (5 6 7) (8 9) (10 11 12 13)
#14	1 (2 (3 4)) (5 6 7) (8 9) (10 11 12 13) 14

を表 2 に、そのスコアを図 4 に示す。また図 5 は#14 で作られた階層構造を図示したものである。(SA とその選択した CA とをまとめて一つの黒点で、リンクは H–リンクのみ「H」の記号をつけて表す。また最終結果に使われていないエージェントは省略する。)

図 4 でステップ#1、#4、#7、#9、#13、#14 は高いスコアになっており、得られた拍節構造 (表 2) は楽譜に対応したものになっている (ただし#9 はさらに 2 拍ずつまとまっている)。これらはちょうど拍位置に対応する箇所である。

逆に拍位置でないステップでは、いくつか楽譜とは異なる解釈の構造が得られている。その中から #8 と #11 における解釈を図 6 に示す。これらの解釈のスコアが高くなったのは、#8 ではそれまでの音長列をほぼ 2 等分、#11 では 3 等分するグルーピングが可能なことによる。これらの解釈は聴き手にとって不自然ではあるものの、ありえない解釈ではない。反面楽譜に対応した解釈のスコアが低くなったのは #8 と#11 がちょうど拍の半分の位置 (裏) にあたり、その断片が音長のばらつきを大きくしたためである。

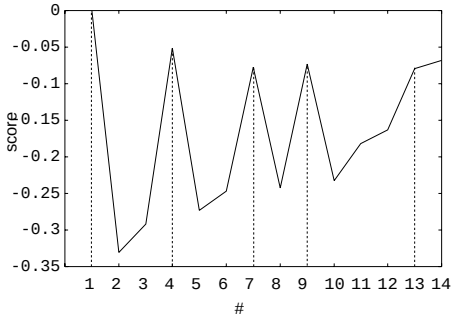

図 4: 各ステップで得られた拍節構造のスコア

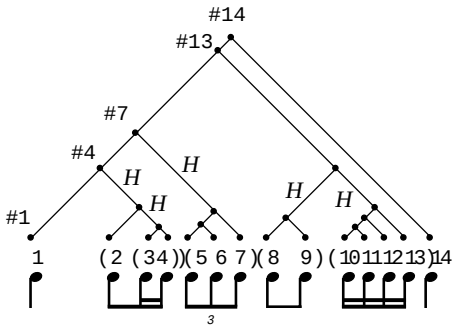

図 5: #14 で得られた階層構造
下段はそれを楽譜で表したものである。

これを解決する方法として、スコアの増減を監視する方法が考えられる。スコアが急に低くなった場合には拍位置でないとみなし、スコアの値がある程度回復するまで出力を抑制すれば拍節の区切りに対応した時点で解釈が得られる。

別な方法として、仮想的な拍位置をあらわす SA を断片の先につけ足し、そこまで含めて解釈することでスコアの低下を抑える方法が考えられる。これは次の拍位置の予測であり、予測処理を加えることでより人間の聴取過程に近いモデルとなることが期待できる。

#8 や #11 で一旦異なる解釈となったにも関わらずそれ以降で妥当な解釈に復帰しているのは以前の処理結果が潜在的には保持され、再び優勢になるためである。また図 5 を見ると#14 で得られた解釈は

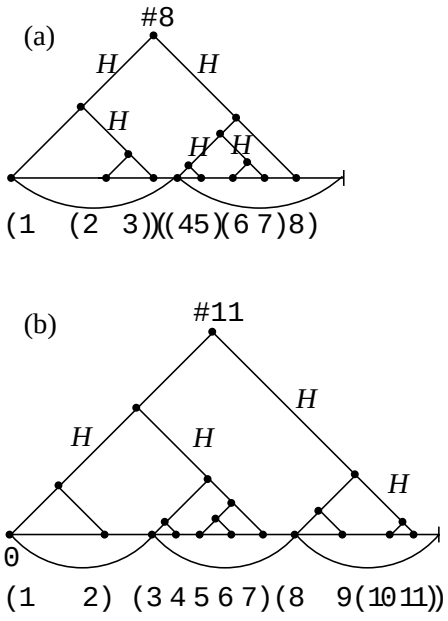

図 6: (a)#8 の階層構造と (b)#11 の階層構造

それまでの拍位置で得られた解釈を含むものとなっている。これは本モデルでは構造が分散的に記憶されていることを示している。

5 おわりに

本稿ではマルチエージェントモデルに基づいた楽曲の構造認識モデルである部分列エージェントモデルを提案し、現時点での実装について報告した。本研究は端緒についたばかりではあるが、実験結果は今後の研究の上で興味深い点をいくつか示している。今後の課題として、評価エージェントを増やし拍節認識以外のグルーピングをおこなうことや、類似性認識エージェントとの協調動作をおこなうことなどがあげられる。

参考文献

[1] Simon, H. A. and Sumner, R. K.: Pattern in Music, *Formal Representation of Human Judgement* (Kleinmuntz, B.(ed.)), Wiley, New York, pp. 219–250 (1968).

[2] 平賀譲: 音楽認知のための知識表現, 音楽と認知 (波多野誼余夫 (編)), 東京大学出版会, 東京, pp. 97–130 (1987).

[3] 平田圭二, 青柳龍也: 音楽理論 GTTM に基づく多声音楽の表現手法と基本演算, 情報処理学会論文誌, Vol. 43, No. 2, pp. 277–286 (2002).

[4] 平田圭二, 平賀譲: GTTM に基づく音楽表現手法再考, 情報処理学会研究報告, Vol. 2002, No. 45, pp. 1–7 (2002). 2002–MUS–45.

[5] Longuet-Higgins, H. C. and Lee, C. S.: The Perception of Musical Rhythms, *Perception*, Vol. 11, pp. 115–128 (1982).

[6] Chung, J. T.: An Agency for the Perception of Musical Beats or If I Only Had a Foot..., Master's thesis, Masachusetts Institute of Technology (1989).

[7] Rosenthal, D. F.: *Machine Rhythm: Computer Emulation of Human Rhythm Perception*, PhD Thesis, Masachusetts Institute of Technology (1992).

[8] Desain, P. and Honing, H.: The Quantization of Musical Time: A Connectionist Approach, *Computer Music Journal*, Vol. 13, No. 3, pp. 56–66 (1989).

[9] Rowe, R.: *Interactive Music Systems: Machine Listening and Composing*, MIT Press (1992).

[10] Minsky, M.: *The Society of Mind*, Simon & Schuster, Inc. (1986). (安西祐一郎訳: 心の社会, 産業図書 (1990)).