マン・マシン協調による採譜システム

半田 伊吹, 木下 智義, 武藤 誠, 坂井 修一, 田中 英彦

東京大学大学院工学系研究科
〒 113-8656 東京都文京区本郷 7-3-1

{handa,kino,muto,sakai,tanaka}@mtl.t.u-tokyo.ac.jp

あらまし:
従来提案されている計算機による採譜システムは、処理全般を計算機に委ねるのが主流であった。しかし、そのようなシステムでは人間が容易に知り得るような情報も計算機では認識が困難である場合もあり、精度の高い採譜は実現しがたかった。筆者らは認識率のより高いシステムを目指し、人と計算機がお互いに得意とする作業を分担し、協調して情報を補完しあう採譜システムを提案する。また、そのようなシステムに必要とされる計算機と人間のインターフェースについての検討を行う。

キーワード: 採譜、マン・マシンシステム

Intelligent music transcription system with human assistance

HANDA Ibuki, KINOSHITA Tomoyoshi, MUTO Makoto,
SAKAI Shuichi and TANAKA Hidehiko

The university of Tokyo
7-3-1 Bunkyo-ku, Tokyo, 113-8656

{handa,kino,muto,sakai,tanaka}@mtl.t.u-tokyo.ac.jp

Abstract:
We have proposed OPTIMA, a system for music scene analysis. The system sometimes misses musical notes. It can be said that complete transcription is difficult for a music transcription system wchich depends on only computational processes. So, we propose a man-machine system so that quality of transcription may inprove. The system contains man-machine interface, and human and machine co-operates in music transcription. We discuss abstract of the system and what kind of tasks can be set on human and computer.

Key Words:
music transcription, man-machine system

1. はじめに

計算機の音楽の世界への応用は様々な試みがなされており、演奏された楽曲の音響信号から楽譜を生成するいわゆる自動採譜に関する研究もなされている。採譜への応用も念頭においた音楽情景分析の処理モデルであるOPTIMA[(1)(2)]は、入力を多角的に解析し、その結果を統合することにより最尤推定像を得ることで精度の高い出力を行うという特徴を備えているが、それでもなお認識精度は実用上十分とは言えず、改善が課題となっている。

OPTIMAの他にも、採譜に挑戦するシステムが提案されているが、実用的な段階には達していないようである。計算機による音楽認知が難しい理由は単に計算量の問題だけではなく計算機の性能が向上すれば解決されるというものではない。音響信号に対してどのようにアプローチすれば精度のよい認識ができるかが明らかにされていないのである。

では、完全に計算機のみによって採譜を行うのではなく、人間と計算機が協調することで採譜の精度の向上は図れないであろうか。計算機では実装が困難な処理であっても人間にとっては容易であることもあり、また逆に人間は不得手であるが計算機が得意とする処理もあるはずである。

本稿では、採譜作業の全てを計算機に委ねることはせず、計算機と人間が強調して採譜を行うシステムを提案し、そのシステムの全体像とマン・マシン間のインターフェース、および人間や計算機に対して無理のない役割分担についての考察を行う。また、計算機へ負担させるべき事項についての基礎的な実験を行った結果の考察を述べる。

2. 計算機と人間の協調による採譜

⟨2·1⟩ 人間にとって容易な処理 一口に人間が採譜を行うと言ってもその人の音楽的な知識や経験の度合によって困難さは著しく異なる。経験豊富な人は少々複雑な楽曲であろうと一度聴いただけでそれを譜面に落とすことができるであろうが、そのようなことができる人は限られている。

一方、特に音楽的経験のない人にとっても、何か新しい楽器がある時刻から鳴り始めればそのことに気付くであろうし、また別のある楽器が周期的にリズムを刻んで鳴っていればそのことに気付くであろう。図1に示すように、人間は楽曲中に含まれる非常に抽象度の高い情報も比較的容易に取得する能力を持っていると考えられる。

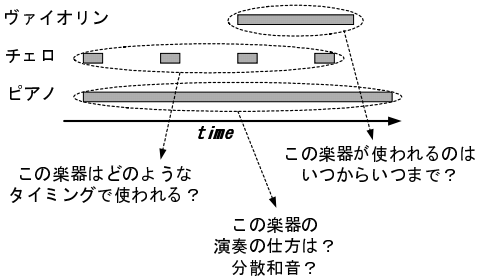

図 1. 人間による楽曲の把握のしかた

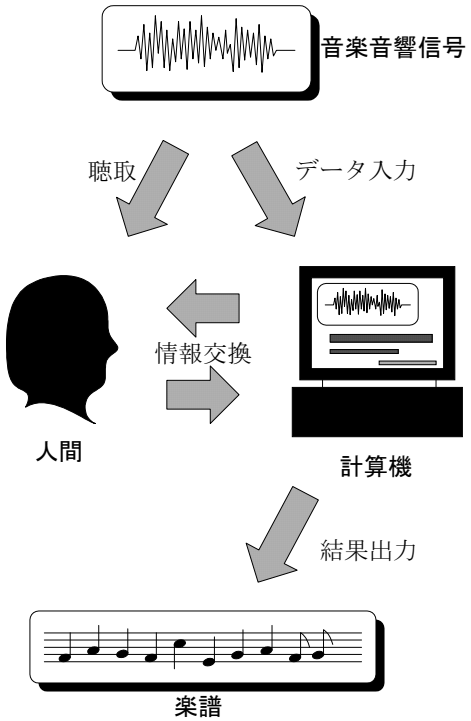

図 2. 提案する採譜システムの概要

⟨2·2⟩ 計算機にとって容易な処理 ある音が単独に与えられたときにそれを他の音に音程的に関係づけずに、それ自身として聴いて、しかもその音の音名を正しく名指す、いわゆる絶対音感を人間が会得するには聴覚的弁別の素質と訓練とが必要とされる。ところが計算機にとっては時間や周波数といった物理量の計測は得意な分野である。一般に楽曲中では複数の音の倍音成分が重なりあってしまうため原因推定が困難になり採譜が難しくなっているが、それでもある微小時間の周波数成分の分布といった人間では知り得ないようなことも計算機ではしっかり把握できる。

〈2·3〉 **システムの概要**　採譜に必要な情報とは、訓練を積んでいない人間の知り得る情報ほどに抽象度が高いものではなく、また物理的特徴ほどに抽象度の低いものでもない。つまり、先に述べた人間と計算機にとって得やすい情報の中間に位置するということができよう。OPTIMA などの音楽情景分析システムは、抽象度の低い物理的特徴量から出発して、より抽象度の高い知識を見出していくという方針を採っている。情景分析という言葉の本義からするとこの方針はもっともであるが、採譜を行うことに重点を置くと別のアプローチが可能であろう。つまり、「情景分析」の結果を採譜に活用してしまうという方法も浮かんでくるのである。

採譜システムを作るということに的を絞れば、人間にとっては比較的に容易に知り得るような情報であっても計算機では抽出困難なものに関しては計算機で不十分な認識処理を行って精度を下げるよりは、人間が計算機を支援して精度の高い認識が可能となったほうが実用的といえよう。

そこで、採譜システムを新たな視点から構築しようと考え計算機と人間の協調による採譜システムを提案する。

提案するシステムの概要を図 2に示す。

まず対象となる音楽音響信号は、人間の耳によって聴取され、計算機には符合化されたデータとして入力される。

楽曲を聴いた人間は、そこに含まれる情報を意識的あるいは無意識的に見出し、ときには情動を揺さぶられることもあるし、はっきり認識しつつも感性には影響をあまり受けない場合もある。楽曲の特徴として

物理的特徴 音楽音響信号の周波数スペクトルの特徴、すなわち音楽音響信号を物理的な音として捉えたときの特徴

音楽理論的特徴 楽曲を楽譜として表現したときに、楽譜上に現れる特徴

感性的特徴 人間が楽曲を聴いたときに受ける印象や感想

の 3 種類を挙げることができる[(3)]。これらは明確な区分があるわけではないが、この分け方に従うと計算機は物理的特徴を正確に認識することを得意とし、人間は音楽理論的特徴を計算機よりはるかに的確に捉えられるのは先に述べた通りである。そこで、計算機と人間はお互いの得た知識をやりとりし、最終的に得た知識をもとに計算機が採譜結果を出力するシステムを提案する。

〈2·4〉 **対象となるユーザと楽曲**　本稿で提案するシステムがどのようなユーザを対象にし、またどのような楽曲に対して採譜を行うものであるか述べておく。現時点ではシステムの提案段階で今後対象が多少変更することは免れないであろうが、最初に明確なビジョンを持って研究を進めるべきであると考えここでまとめておく。

まずユーザについては、音楽の専門家ではないが義務教育程度の音楽の知識を有する人を対象とする。

採譜の対象となる楽曲については、唱歌に簡単な伴奏をつけたもの (音声は対象としていないので楽器に置き替えて演奏したもの) を想定している。使用する楽器は基音に対して非整数倍の成分をほとんど持たないものとする。

3. インターフェース

ここでは計算機と人間がどのようなやりとりをどのようにするかについて検討する。

まず大前提となることは、人間の行うべき作業が極端に繁雑になってはいけないということである。人間だけで採譜を行う場合よりも面倒なものになっては本末顛倒である。そこで、システムには、

- 人間のみが採譜を行う場合より、より多くの情報が得られる
- 人間のみが採譜を行う場合より少ない作業量で採譜が行える

といった仕様が要求される。ここでは特に後者を満たすことを考えることとする。

例えば図 3のような楽曲を認識する際に、計算機上では図 4のように時間、音高で区切った空間を埋める問題として処理しているとする。このとき、計算機がマトリックスを示してあとは人間が穴を埋めてくれれば計算機への負担は少ないが、人間にとっての負担は大きい。音楽の専門的知識がない人間でも容易にできることといえば、音が鳴っているか否かの判断や音高の変化のおおまかな判断くらいであろう。そこで、図 5に示すように、計算機は発音時刻の候補を示し、人間はそこに音の有無および音高について一つ前の発音とのおおまかな相対関係 (つまり上行したか下行したか) だけを入力するようなシステムが要求仕様を満たすと考えられる。計算機側が担う発音時刻の特定は、抽象度の低い情報から得ることの比較的容易な問題であると考えられ、次節で簡単な検証を行う。

人間が入力する情報のうち、単音が鳴っているの否かの情報は非常に重要である。このように計算機にとっては大変有用な情報を人間が簡単に入力できるようなインターフェー

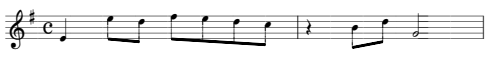

図 3.　譜例 a

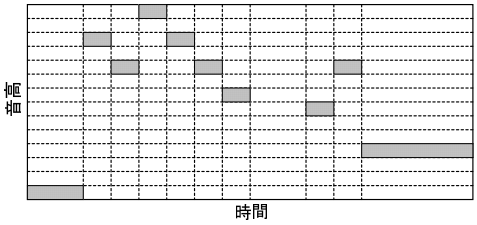

図 4.　抽象化された楽曲

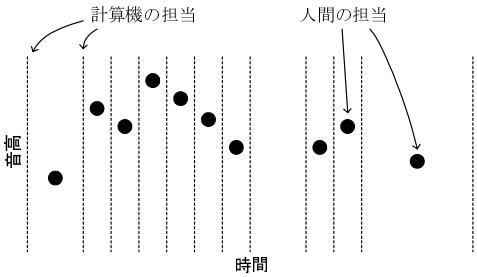

図 5.　計算機の出力情報と人間の入力情報

スが望まれる。

4.　計算機上で実装可能な技術

本節では、前節で例示したインターフェース機構の基礎となる計算機側での処理について述べる。ユーザ (人間) が簡単に石を置くような感覚で音の有無や上行・下行の情報を入れられるためには、発音の起こる時刻を計算機側で認識しはっきり線引きして提示することが要求される。線引きがはっきりしていれば単純な穴埋め問題に帰着し、ユーザが与えるヒントを受けて再び計算機の出番となりその問題を解くことになる。

⟨4·1⟩　発音時刻の抽出　　発音時刻を抽出するためには、音響信号に Q 値が一定のバンドパスフィルタバンクを介して得たスペクトログラムを用いることにする。時間と周波数は不確定性原理の関係にあり、両者の分解能を同時に高めることは不可能である。時刻の検出を目的としているので、ここで用いるスペクトログラムは Q 値を低くして周波数分解能が悪く時間分解能が良いものとする。

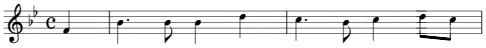

図 6.　譜例 b

時間を t、周波数を fとして、時間–周波数解析の結果求められたパワー (スペクトログラム) を $P(t,f)$ とする。発音時刻検出のための関数 Q を以下のように定義する。

$$Q(t) = \sum_f \delta(t,f)\,(P(t,f) - P(t-1,f)) \quad \cdot(1)$$

ただし

$$\delta(t,f) = \begin{cases} 1, & P(t,f) - P(t-1,f) > 0 \\ 0, & P(t,f) - P(t-1,f) \leq 0 \end{cases} \quad (2)$$

である。$Q(t)$ を参照して、以下の条件を満たす時刻 t を発音時刻とする。

$$\begin{cases} Q(t) - Q(t-1) > 0 \\ Q(t+1) - Q(t) \leq 0 \\ Q(t) > Q_0 \end{cases} \quad \cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot(3)$$

このようにして発音時刻を検出することにし、実際に簡単な例に対して予備実験を行った。まず、図 6に示す譜例をピアノで演奏した入力に対して図 7のようなスペクトログラムを求める。次にこれをもとに式 (1)〜(3) によって発音時刻を求める。得られた結果は表 1に示す。1 分あたり四分音符 120 拍の速度で曲は演奏されているので、正確に発音時刻の検出が行われたことがうかがえる (なお、本来の発音時刻より少しずれた結果になっているのは誤検出ではなく、元の演奏に意図的に発音時間のゆらぎを加えているためである)。

この実験では正解の得られる閾値 Q_0の設定範囲も広く実用に適する印象を受けたが、複数楽器の合奏にこの処理を施して発音時刻を検出しようとしたところ、閾値の設定のしかたによってとりこぼしがあったりノイズを拾ってしまったりといった不都合が生じてしまった。式 (1) のように評価指標を 1 次元に落としてしまっているためにこのような困難が生じているのでないかと考えており、検討を要する。

⟨4·2⟩　音源・音高の同定　　ある発音時刻に着目して、その時刻にどんな楽器がどんな音高で鳴ったのか同定できれば採譜できたことになる。図 8に示すように楽器によってその周波数成分 (基音のパワーに対する倍音のパワーや時間変化) は異なる。使用が想定される楽器の周波数成分の特徴のテンプレートを

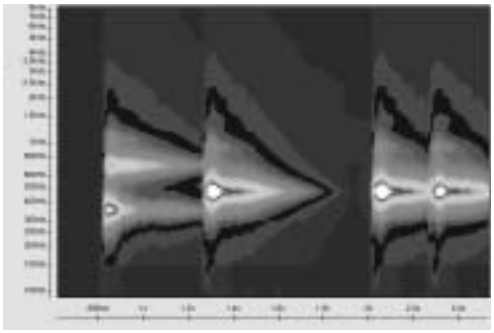

図 7.　スペクトログラム (Q 値の小さい場合)

表 1.　発音時刻

取得番号	時刻 (ms)
1	0
2	446
3	1186
4	1469
5	1937
6	2493
7	3200
8	3450
9	3951
10	4201

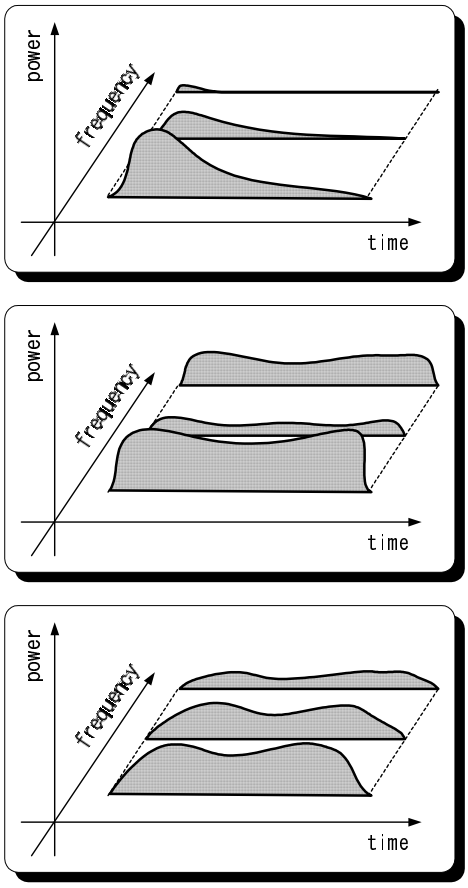

図 8.　周波数成分の特徴の例

予め用意しておき、実際の周波数成分と比較することで音源・音高が同定できるであろう。ここで用いる周波数成分は、図 7のような周波数分解能の低いスペクトログラムではなく、図 9のような高い Q 値のフィルタによって得られた周波数分解能の高いスペクトログラムも用いて得られると考えている。

発音時刻においては、図 10に示すように全ての周波数成分が新規に生起するのとは限らず以前の周波数成分が継続する場合もある。このことに注意しなければならないが、発音時刻を予め抽出することによって周波数成分が生起したか否かを判断する時刻が限定され、計算量が少なくかつ確度の高い成分抽出が可能となるであろう。

5.　おわりに

本稿では、計算機だけでは困難な採譜処理に人間の助けを加えて構築する採譜システムを提案した。この手法により認識精度が従来より高い採譜システムができるものと考えられる。

このように機能の面で優れていても、完全に計算機に処理を委ねたシステムに比べ、認知科学や計算機科学の観点からは消極的でつまらないものという印象を与えかねない。しかし、どのような情報を得ると認識精度が向上するのかという見識が、本システムの実装を行う上で得られるであろうから、結果として学問的にも意義のある研究である。また、物理的な特徴量と、人間が言語をもって辞す

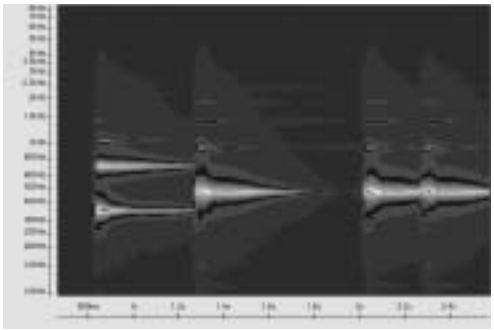

図 9.　スペクトログラム (Q 値の大きい場合)

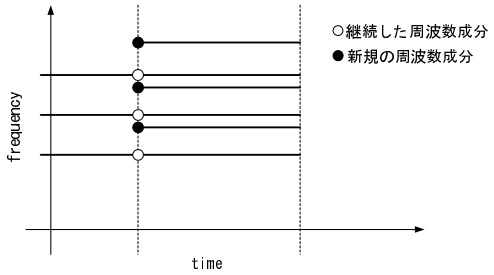

図 10.　新規の発音と継続音

るような高度な情報を統合する機構を採譜システムに導入することは新奇であると考えている。

今後は計算機による発音時刻検出の確度向上と、発音時刻が与えられたときの音源・音高同定モジュールの開発、および操作性の良いインターフェースの模索を進めていく予定である。

文　　献

(1) 柏野邦夫, 中臺一博, 木下智義, 田中英彦: 「音楽情景分析の処理モデル OPTIMA における単音の認識」, 電子情報通信学会論文誌, Vol.J79-DII, No.11, pp. 1751–1761, 1996

(2) 柏野邦夫, 木下智義, 中臺一博, 田中英彦: 「音楽情景分析の処理モデル OPTIMA における和音の認識」, 電子情報通信学会論文誌, Vol.J79-DII, No.11, pp. 1762–1770, 1996

(3) 武藤誠, 木下智義, 半田伊吹, 坂井修一, 田中英彦: 「音楽音響信号からの楽曲の感性的特徴の抽出」, 情報処理学会平成 11 年後期全国大会講演論文集 (2), 4G-6, pp. 2-11–2-12, 1999