音楽情報科学 30－2
（1999．5．29）

音楽認識の評価方法に関する考察

鈴木 啓高　中西 正和

慶應義塾大学大学院 理工学研究科 計算機科学専攻
{hiro,czl}@nak.ics.keio.ac.jp

本稿で音楽認識処理の出力結果，特にビートトラッキングの出力結果の評価方法を提案した．従来までの音楽認識システムの研究において，システムの出力結果の評価方法は確立されていない．対象となる音楽自身が絶対的な評価軸を持たず，また音楽特有の曖昧性のために，完全に誤りとは言えないような誤りが存在するため，客観的な評価を困難にしている．したがって，その多くはある特定の人間，あるいは複数の人間による主観に基づいた評価をシステムの出力に対する評価としていた．本稿では，従来手法で用いられていた正解と出力との時間的ずれに加えて，正解と出力それぞれのデータの時間的間隔を評価を行なうための要素として取り扱うことで，より柔軟に評価を行なうことが可能となった．

A study of way to evaluate the result of music recognition

Hirotaka SUZUKI　Masakazu NAKANISHI

Department of Computer Science
Graduate School of Science and Technology
Keio University
3-14-1, Hiyosi, Kouhoku, Yokohama 223, Japan

In this paper, we propose a new method to evaluate the result of a music recognition system. Due to a vagueness of music itself, no definite method of evaluating such systems has been establicshed for the present. A time lag between the correct beat and the output beat of a system is used as a basis of their evaluation. By add a transition of every two consecutive betas in both the correct data and the output data, it become possible to deal with a vagueness of music. As a result of some experiments, a similar value to a feeling of a human being was recognised.

1 はじめに

近年の音楽情報処理の分野において，一般の音楽音響信号を対象とした認識システムの研究が盛んに報告されている．しかし，もともと絶対的な評価軸を持たない音楽を対象としているため，その評価は困難であり，決定的な評価方法というのは未だに確立されていないといえる．現状では，システムの出力をある程度の人数の人間がそれぞれの主観で評価を行ない，それらをまとめてシステム全体の評価としているのが主であるといえる．今後さらに一般の音楽音響信号を対象とした認識システムの研究を行なっていく上で，客観的かつ定量的な評価方法の確立は急務であると考えられる．本稿では，音楽認識の研究の中でも特にビートトラッキングシステムに焦点をあて，その評価方法について検討する．

2 従来手法

後藤が文献 [3] で述べているように，これまでビートトラッキングのシステムの評価尺度に関する議論はほとんどなされていない．様々なシステムそれぞれに特化した評価手法は報告されているが，より汎用的な手法は未だに確立されていない．

後藤は階層的なビート構造の各レベルにおいてトラッキングの精度を分析でき，典型的な誤認識[1] を同定できる評価方法を提案した [3]．

後藤は正解時刻列と出力時刻列とから，まず最も近いデータ同士をペアとし，次にそれぞれのペアの時刻のずれから 4 分音符，2 分音符，小節レベルの各レベルについての評価を行なった．それぞれのレベルについて，正しくビートトラッキング出来た期間，ペアの時刻のずれの平均，標準偏差，最大を求め，同時にテンポ誤り，位相誤りを表すフラグを用意した．

[1]ここでの誤認識とは正解ではないが完全に誤りとはいえない出力を指す．例えばビートの間隔を実際の半分として出力した場合などがこれにあたる．

3 議論

本章では計算機上のシステムの出力を評価するにあたって要求される事項とその問題点を検討する．

3.1 評価の方針

ここで検討する評価方法としては「システム[2] からの出力全体に対する評価値を出力」するものとする (図 1)．また理想的な評価方法の特徴として以下の 2 点を挙げる．

- 人間がシステムの出力を聴いた時に持つ心理的感覚を可能な限り反映した数値を出力する
- 必要となる閾値は出来るだけ少なくし，機械的に処理できるものとする

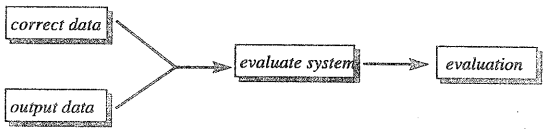

図 1: 評価システム概要

音楽に絶対的な評価軸が存在しないという理由から，従来手法の評価においては人間の主観による評価が主流であった．しかし統計的に有効な評価とするためには非常に多くの異なる人間による評価が必要となる．これは実験を行ない評価を行なう際に毎回大勢の人間を集める必要があるという意味であまり現実的でない．また，評価を行なう人間の選択方法の面からも客観的な評価方法としては用いにくい．大勢の人間による評価の平均値に近い値を出力する

[2]一般の音楽音響信号を入力とし，その拍の位置を認識し出力するものを指す．取り扱う信号は時間軸に沿ったデータ列とする．

評価システムを構築することで，より人間の感覚に近くかつ客観的な評価を行なうことが可能であるといえる．

また，より客観性を増すためには評価方法自体から可能な限り人間の主観を除く必要がある．そのために人間の経験と勘が色濃く出るような閾値の利用を極力避ける必要がある．

3.2 評価に伴う問題点

次に先に述べた要求を満たす評価の手法について検討する．正解の作成が困難であることに加えてビートトラッキングの評価を困難なものにしていると考えられる主な要因として以下の3つを挙げることが出来る．

- 正解の拍と認識システムの出力する拍とか必ずしも1対1に対応するとは限らない
- 正解の拍とはずれているが，人間が聴いた場合に必ずしも誤りとはいえない認識システムの出力が存在する
- 認識システムにおける拍認識のレベル[3]の同定が困難である

正解と出力との比較を行なう際には，それぞれの時系列データの比較を行なうことになる．しかし，その際必ずしも各データが1対1対応になるとは限らない．例えば，最も近いデータ同士をペアとして定義した場合，1つの正解データに対して複数の出力データが同時にペアとして定義される可能性がある(図2-case1)．各ペア同士のずれを評価の基準として評価値を求める手法において，このようなペアの重複を考慮にいれないと正しく評価を行なうことが出来ず，実際よりも高い評価値を出力することになる．

[3]認識システムの出力する拍が何分音符単位のものなのかをさす

また逆に，ペアとして関連づけられた出力データを持たない正解データが存在するような場合もある(図2-case2)．この場合も先に挙げた例と同様に，ペアとなる出力データを持たない正解データの取り扱いが問題となる．

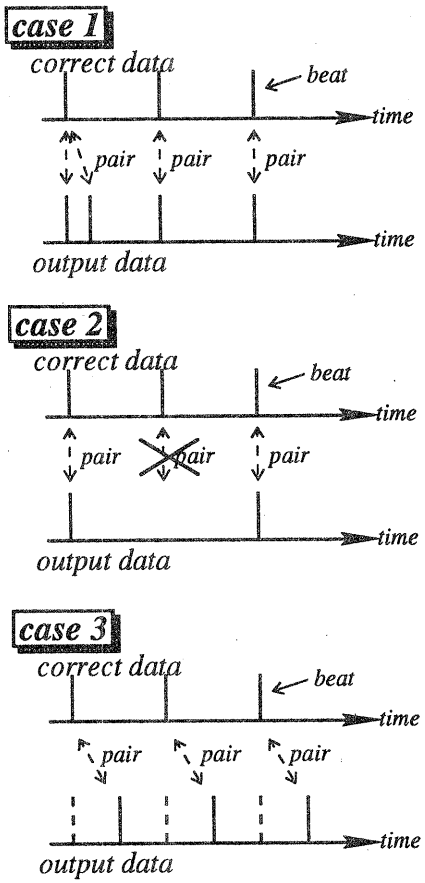

図 2: 評価が困難な例

ビートトラッキングシステムの出力結果を評価する場合，評価の対象は時系列データとなる．したがって評価は時間軸に沿ってデータを走査していき，正解とのずれを何らかの処理を通して評価結果へと結びつけていくことになる．人間がビートトラッキン

グシステムの出力結果を評価する場合, 正解と出力とがずれていても単純に誤りとは感じない場合がある. 例えば, 全て裏打ちで入っている場合などがこれにあたる (図 2-case3). このような場合には, 他のずれとは別にデータを取り扱う必要がある.

ここで例えば半拍ずれとなった場合, それが全くの偶然によるものなのか, あるいはシステムに実装されている音楽的知識によるものなのかを正しく判断し評価を行なうことが必要となる. しかし, 偶然あるいは必然のどちらである場合でもそれがシステムの出力であるのであれば平等に評価の対象となると考え, 偶然か必然かの考察は評価をする時点では必要ないということも出来る.

ビートトラッキングシステムの出力結果から入力の音楽のテンポを求める場合にはシステムの認識レベルも併せて評価する必要がある. しかし, これも音楽が絶対的な評価軸を持たないという理由から実現は困難である.

3.3 評価基準について

ある入力に対して何らかの処理を行ないデータを出力するシステムを評価するにあたって, まず, 出力データの正当性を計るための基準 (正解データ列) が必要となる. 特に対象を音楽としたシステムの場合, この正解データの取得は非常に困難な作業になるといえる. ここでもやはり, 簡単な手法としては,

ある個人の主観で正解を作成
⇓
得られた正解を十分な人数の人間が評価

この過程を繰り返し, 十分な正当性を示すというのが考えられる. しかし, 処理の対象となる音楽が絶対的な評価軸を持たないために, 全ての処理の過程において曖昧性が残ってしまう. そのため, 全く機械的な処理で正解を作成することが可能なのが一番理想ではあるが, それは実現困難である.

そこで, 多少なりとも客観性を高めるために楽譜情報を用いる手法が考えられる. これは, 認識対象となる音楽の楽譜情報を用意し, これを元に正解となる拍を決定するものである. しかし, これもやはり楽譜情報の正当性という面で先に挙げた手法と同じ問題が発生する. いかにしてもっともらしい正解を作成するかが課題であるといえる.

4 評価方法の提案

本章で, より客観的かつ機械的な処理による評価方法を提案する. ここで認識の対象となる音楽は, 全体を通してテンポがほぼ一定であるものに限定する.

4.1 手法

従来手法と同様に, 時系列データである正解データ列と出力データ列とのずれをもとに評価を行なう. はじめに, ずれを求めるためのペアを作成する. 正解データ列を C_n とし, システムの出力データ列を O_m とした時に式 (1) を満たすもの同士をペアとする.

$$\begin{cases} |\Delta T| \leq (C_{n+1} - C_n)/2 & (\Delta T \leq 0) \\ |\Delta T| \leq (C_n - C_{n-1})/2 & (\Delta T > 0) \end{cases} \quad (1)$$

ただし, ここで ΔT は以下のように定める.

$$\Delta T = C_n - O_m$$

ここで, あるひとつの正解データに対して複数の出力データがペアとして作成された場合も全て異なるペアとして取り扱う. したがって, 作成されるペアの数はシステムによる出力データ列のデータ数と同じになる.

また, 正解データは認識対象となる音楽の楽譜情報をもとに作成したものを用いる. ここで正解データ列は認識レベルが 8 分音符, 4 分音符, 2 分音符小節レベルの各レベルのものの 4 種類を作成する.

作成されたペア同士のずれを元に評価を行なっていくのだが, 本稿で提案する手法では先に述べた「ひとつの正解データに対して複数の出力データがペアとして作成された場合」に対処するために「ペア同士のずれ」に加えて「正解データ列の連続する 2 つのデータの時間隔と出力データ列の連続する 2 つのデータの時間隔の変化」を評価を行なうための要素として取り扱う.

これは正解データ列とは異なっていても, ビートトラッキングの出力がある一定の間隔を保っていれば, 人間が聞いた場合にそれを完全に誤りとは感じない場合があることによる. 例えば, 認識を行なう区間の始点から終点まで裏打ちを行なった場合がこれに相当する. 「連続する 2 つのデータの時間隔」を評価軸として用いることで, これまで困難であった「完全に誤りとは言えない結果」を評価することが可能となる. また, ペア作成において, いずれの出力データ列ともペアとして関連づけられなかった正解データが存在するような場合にも正しく対処することが可能となる.

「ペア同士のずれ」および「連続する 2 つのデータの時間隔」を全てのペアおよびデータに関して求め, これを元にビートトラッキングシステムの出力に対する評価値を決定する.

従来手法が, 正しく拍が認識出来ている区間に対してポイントを与えていくという意味で加点法であったのに対し, 本稿で提案する手法では完全に正しく認識出来ている場合を 1 とし, そこから誤認識の分を引いていく減点法を用いる. 減点法を用いることで 1 つの正解データに対して複数の出力データがペアとして作成された場合に正しく対処することが可能となる.

まず「ペア同士のずれ」の分布を参照し, その広がり具合を減点分とする. 全体で N ペア作成され, その中に含まれるあるペア同士のずれを y_n, ペアの内の正解データの時刻を C_n としたとき, 評価値 $Point$ を式 (2) で定義する.

$$
\begin{aligned}
Point &= \frac{N - \sum_N (1 - \sum_{pair} point(n))}{N} \\
&= \frac{\sum_{pair} point(n)}{N} \qquad (2) \\
point(n) &= \exp(-\frac{(\frac{6y_n}{beat(n)})^2}{2}) \qquad (3) \\
beat(n) &= (C_n - C_{n-1})/2 \qquad (4)
\end{aligned}
$$

次に「連続する 2 つのデータの時間隔」の変化を参照する. 小節, 2 分音符, 4 分音符, 8 分音符の各レベルの時間隔の変化のうち, 出力データの時間隔の変化と最も距離[4]が近いものを評価に用いる. 出力列および正解列の時間隔の変化, およびそのずれの広がり具合を減点分とする. ペア同士のずれがなく, かつ時間隔の変化が等しければ正解列と出力列は一致することになる. ここで, いずれの出力列ともペアとして作成されなかった正解列が存在する場合でもデータの時間隔の変化とそのずれが等しい場合が「完全に誤りとはいえない誤り」に相当する.

4.2 実験

本稿で提案した評価方法を用いて実際のビートトラッキングシステムの評価実験行なった. 実験には文献 [5] で提案されたビートトラッキングシステムを用いた. 時系列である入力データとシステムの出力データとを用いて評価を行なった. 正解データ列は入力の音楽音響信号の楽譜情報から人間が作成した.

[4] 各時刻における時間隔の差の総和を距離とする

表 1: ペア同士のずれによる評価 (%)

小節	2 分音符	4 分音符	8 分音符
41.11	88.23	88.23	82.35

4.3 結果

実験に用いた曲は前半では 4 分の裏を, そして後半では 2 分でビートトラッキングを行なった. ペア同士のずれだけで評価をした場合の結果を表 (1) に示す. これらの値からシステムの出力に対する評価値を決定した場合, 高々88.23%となる. しかし, これを実際に人間が聞いた場合の感覚的な評価はこれよりも高い値となった. 正解と出力の各データ列の時間隔の変化も合わせて評価を行なった結果 93.84%という評価値を得た. これは, より人間の感覚に近い値であるといえる.

5 おわりに

本稿で我々は音楽認識システム, 特にビートトラッキングシステムの出力結果に対する評価を行なう際の問題点を示し, 新しい評価方法を提案した. ペア同士のずれだけでなく, 正解データ列と出力データ列の各データ間の間隔も評価を行なう際の要素に含めることで「正解とは一致しないが完全に誤りとはいえない」出力を評価することが可能となる. しかし, ビートトラッキングを行なう区間において徐々に正しく拍を認識していくような出力結果の場合の取り扱いなどの問題もある. 今後はさらに客観的かつ汎用的な評価方法の検討を行ない, 人間が音楽を聞いた時に抱く感覚を可能な限り反映することの出来る評価システムを構築したいと考えている. また本稿で提案した評価方法を単に評価を行なうためだけでなく, さらに様々なシステムへと応用していくことを検討する予定である.

参考文献

[1] David Rosenthal. Emulation of Human Rhythm Perception. *Computer Music Journal*, Vol. 16, No. 1, Spring, 1992.

[2] Large E. W. Modeling Beat Perception with a Nonlinear Oscillator, *In Proc. of the 18th annual Conference Cognitive Science Society*, pp.420–425, 1996.

[3] 後藤真孝. 拍節認識 (ビートトラッキング), コンピュータと音楽の世界, pp.100–116, 共立出版, 1998.

[4] 後藤靖宏, 阿部純一. "リズム" 認知過程のモデル化:" メトリカル・ユニット階層化モデル" と" 内的クロック生成モデル" の比較と今後の方向性, 日本認知科学会 第 12 回大会, pp.152–153, 1995.

[5] 鈴木啓高. 音楽音響信号に対するビートトラッキングのための拍認識, インタラクション '99 , pp.133–134, 1999.