音楽ロボットとの合奏ための信頼度を用いたビートトラッキングの結合手法

○糸原 達彦,　奥乃 博　(京都大学大学院　情報学研究科)

1.　はじめに

本論文では人とロボットの合奏のための信頼度を用いたビートトラッキングの結合手法について報告する. 合奏において, 音楽ロボットはビートトラッキングと呼ばれるテンポと拍時刻を推定する手法が必要になる. 従来, 様々なビートトラッキングが開発され, ロボットとの合奏に適用した例が報告されている. これらのビートトラッキングにおいて, テンポや拍時刻の推定には独自の仮定が設けられており, それぞれに利点, 欠点が存在することが明らかになっている. そのため, 現在では合奏したい曲目によってビートトラッキングを適切に選ぶ必要が発生し, またそれが原因でロボットとのスムーズなインタラクションが妨げられている.

本研究では信頼度を各ビートトラッキングごとに計算することで, 各手法を結合しよりよいビートトラッキングを達成することを目的とする. 信頼度は入力信号スペクトログラムの一小節長の 2 フレーム間における自己相関を元に計算される. この値は入力信号及び推定拍時刻のみから計算されるので, あらゆるビートトラッキング手法に適応することが可能である. 並列に計算されたビートトラッキング結果の中から適切なものを逐次的に選び出すことが可能になる.

本稿では村田らと糸原らの手法の結合を行なう. この結合の目的は, 演奏楽曲がオンビート音楽かオフビート音楽かによらずに一定のパフォーマンスを得ることである. 実験として結合前のビートトラッキング手法との比較及び評価を行い, オンビート音楽及びオンビートオフビート混合音楽での精度向上を確認した.

2.　研究背景

本研究の目標は人とロボットとの音楽合奏を実現することである. 人とロボットのインタラクションの確立は人とロボットの共生のために非常に重要な要素であり, 音楽ロボットとの同期の実現は, 言語のような意思疎通における障壁が少ないという音楽の特徴があるため, その共生を促進することが期待されるものである.

音楽ロボットが合奏するための必要条件は 2 つに分けられる. 一つは楽器を弾く能力, もう一つは人の演奏に対する同期である. 楽器を弾くロボットの開発は多数行われてきている. Solis らは肺の構造を模倣したフルート演奏ロボットの開発を行い, 息のコントロールひいてはフルート演奏技術の向上を報告した [1]. 水本らはロボット用テルミン演奏モデルを開発した [2]. このモデルはロボットに依存する部分と依存しない部分の分離の観点から開発されており, このモデルにより様々なロボットでテルミンを演奏させることが可能になった.

音楽ロボットとの合奏の研究は盛んに行われてきた.

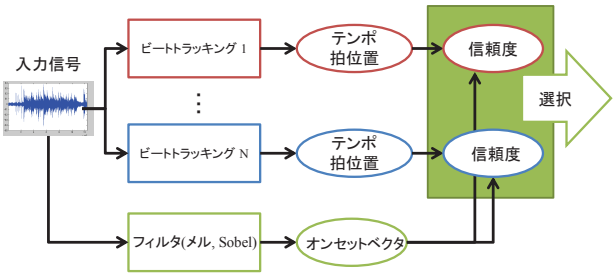

図 1 結合されたビートトラッキングの概要図.

Weinberg らは共演者のパーカッション演奏の模倣及び同期をするロボットを実現した [3]. 水本らはテルミンロボットと人のパーカッション, フルートの演奏者での三重奏を報告している [4]. また, Hoffman らはマリンバロボットとの即興セッションを実際の観客の前で行なった例を報告している [5].

これらのロボットが演奏者と同期するためには, 共演者の演奏のテンポ及び拍時刻を推定する必要がある. 拍時刻とは一小節を 4 分音符間隔で等分した時の時刻を指すものとする. この推定を行なうのがビートトラッキングである.

音楽ロボットのためのビートトラッキングが解決すべき問題は, (1) 環境ノイズの混入, (2) テンポの流動性, (3) ビートの複雑さ, の 3 つである. しかし, これらすべてを同時に解決した手法は今までには存在せず, ロボットとの合奏において, 曲目によってビートトラッキング手法を交換する必要があった. 例えば, 村田らは STPM(Spectro-Temporal Pattern Matching) という人のテンポ変化やロボットノイズに頑健なビートトラッキング手法を開発し, ロボットが音楽合わせてステップを踏む実験を報告した [6]. しかし, この手法はオンビート音楽を仮定した手法であり, オフビート音楽では追従結果が低いという問題がある.

糸原らはギター演奏者に対する視聴覚統合ビートトラッキングを報告した [7]. この手法は, (1) だけでなく, (2) と (3) のトレードオフを同時に解決することを目的としており, その達成のために, STPM から得られる音声特徴量に加え, ビートパターンと相関の深いギター奏者の手の動きを入力とし, それらをパーティクルフィルタで統合している. この手法は, オフビート音楽では村田らの手法に比べ高い追従結果を報告しているが, オンビート音楽では村田らの手法の方がより高い精度で追従している. またこの手法はギターに限定されており, ギター以外の楽器では高い水準は期待できない.

他にも, マリンバのマレット追従によりビートトラッキング [8] や, フルートのジェスチャー認識を交えた合奏 [9] などが報告されているが, これらは楽器への依存

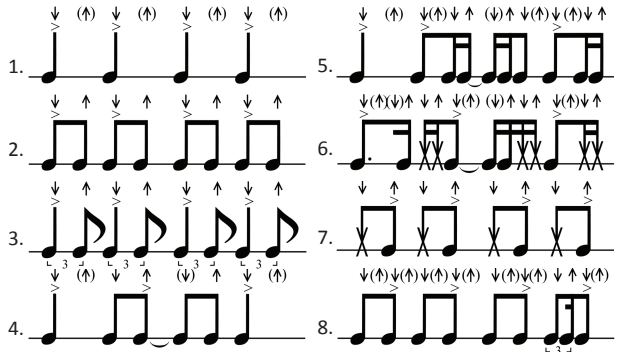

図 2 代表的なギターのビートパターン．× は素早く音をミュートすることで打撃音のを出す奏法 (カッティング) を，> はアクセントを，矢印は手の運動方向を，括弧つきの矢印は空振りを表す．

性が高く，糸原らの手法同様，楽器の有無により手法を変える必要がある．この交換の必要性は特にセッションのような，合奏中に合奏のリズムを決定している人が変化する場合において問題であり，人とロボットのスムーズな合奏を阻害するものである．

3. 本手法で扱う合奏の仮定と問題

3·1 合奏の設定

合奏の構成は，メロディー担当ロボット 1 体と伴奏担当の人のギター奏者 1 人であるとする．ギターを採用する理由は，ギターの演奏容易さや伴奏楽器としての一般性からくる，ロボットとの演奏機会の増加の狙いがある．問題の簡略化のために 4 分の 4 拍子の音楽を仮定する．また，伴奏担当は同じビートパターンで演奏を繰り返すものとする．この仮定が成立しない場合は，テンポが変わらないという前提でない限り，人も拍追従ができなくなるためである．

演奏の初めに，共演者とのタイミングやテンポを合わせるために “カウント” を行うものとする．これは主に声やギターの打撃音で行われる．4 分の 4 拍子の音楽の仮定に基づき，カウントの数を 4 と固定する．本手法では，楽譜は用いない．理由は，(1) インターネット上などで公開されている多くのの楽譜にリズム譜が記載されていないこと，(2) 本手法の目標が即興演奏であること，である．また，ロボットは自身に内蔵された，もしくはロボットの近くにおかれたマイクで音を検出するとする．その理由は，合奏の規模を拡大する時にマイクを購入するなどのコストが発生しないようにするためである．

3·2 本手法のビートトラッキングの入出力と問題

音楽ロボットのためのビートトラッキングが解決すべき問題は，(1) 環境ノイズの混入，(2) テンポの流動性，(3) ビートの複雑さ，の 3 つである．問題 (1) は，入力の音楽信号にロボットのファンノイズやエアコンなどの周囲物の音が混入することを指す．従来のビートトラッキング研究ではロボットのファンノイズなどに言及したものは少ない．しかし，ビートトラッキングにおいて，ロボットから発生する音の影響は大きく，ノイズに頑健な音響特徴量を使う必要がある．

問題 (2) は，合奏相手として，プロの演奏家を想定してないために発生する．誰でも参加できる合奏を目指すためには，この問題は解決されなければならない．

問題 (3) はオフビート (裏拍) 音楽などにおけるシンコペーションなどによく見られる．ここで裏拍及び表拍は，一小節を偶数個に等分したときのそれぞれ偶数番目，奇数番目の拍のことを指し，シンコペーションとは偶数番目のとその次の奇数番目の拍との連結のようなビートパターンを指す．図 2 に，代表的なビートパターンを示す．パターン 1,2 はパターンの基礎となるもので，3 はその 3 連符が混ざったものである．これらのアクセントはすべて表拍に置かれる一方，それ以外は裏拍アクセントを含んでいる．パターン 4 がシンコペーションの代表例である．パターン 7,8 のアクセントは裏拍にのみ置かれている．以上より，アクセント位置に対する頑健性は重要である．

特に 問題 (2) と (3) の同時解決が難しい理由は，両者がトレードオフの関係にあることである．例えば，テンポの流動性に敏感になりすぎるとビートパターンの複雑さをテンポ流動性として吸収してしまう．逆にビートパターンに頑健にしようとすぎるとテンポ変動を無視する恐れがある．

4. 信頼度

汎用的に使用可能な信頼度はビートトラッキング手法によらず同じ方法で計算されなければならない．例えば，体重と身長といった異なる値の比較をすることはできないからである．以上より，信頼度の計算に用いる情報は，入力の音声信号とビートトラッキングの出力であるテンポ及び拍時刻であるとする．

本手法では入力の音声信号をオンセットベクタに変換して扱う．オンセットベクタとは入力信号のスペクトログラムにメルスケールフィルタと Sobel フィルタをかけたものである．それぞれのフィルタの目的は，計算量の削減及び定常ノイズの抑圧である．以下でその導出を説明する．

入力となる音楽信号を 44.1[kHz]，16[bit] で同期してサンプリングしたのち，窓長 4,096[pt]，シフト長 512[pt] で短時間フーリエ変換 (STFT) を用いた周波数解析を行う．得られたスペクトルにメルフィルタバンクを適用し，周波数の次元数を削減した．本稿では 15 次元にした．得られたメルスケールでのパワースペクトルを $p_{mel}(t,f)$ とする．f はメル周波数軸での周波数インデックスを表す．t フレーム目のスペクトログラムに対し，エッジ強調をするために Sobel フィルタを適用し，負の部分を 0 としたものをオンセットベクトル $d(t,f)$ と定義する．$d(t,f)$ は以下の式で導出される．

$$d(t,f) = \begin{cases} p_{\mathrm{sobel}}(t,f) & \text{if } p_{\mathrm{sobel}}(t,f) > 0, \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$\begin{aligned} p_{\mathrm{sobel}}(t,f) = &-p_{\mathrm{mel}}(t-1,f+1) + p_{\mathrm{mel}}(t+1,f+1) \\ &-p_{\mathrm{mel}}(t-1,f-1) + p_{\mathrm{mel}}(t+1,f-1) \\ &-2p_{\mathrm{mel}}(t-1,f) + 2p_{\mathrm{mel}}(t+1,f). \end{aligned} \quad (2)$$

ただし，p_{sobel} は Sobel フィルタの出力である．

ここで, あるビートトラッキング X について, ビートベクタ $q_X(n,m,f)$ を導入する. ビートベクタはオンセットベクタの拍位置間を m_q 個のベクタに量子化したものであるとし, n,m,f はそれぞれ検出拍インデックス, 拍間インデックス, オンセットベクタの周波数ビンである (式 3). 式中の $t_{X,n}$ はビートトラッキング X が検出した n 個目の検出拍の時刻であるとする. ビートベクタの要素の詳細な定義を式 3 に示す.

$$d([t_{X,n}, t_{X,n+1}-1], f) \Rightarrow q_X(n, [0,1,\ldots,m_q-1], f) \quad (3)$$

$$q_X(n,m,f) = \frac{1}{t_{X,n+1}-t_{X,n}} \sum_{i=m\Delta t_X}^{(m+1)\Delta t_X} d(t_{X,n}+i, f) \quad (4)$$

$$\Delta t_X = \frac{t_{X,n+1}-t_{X,n}}{m_q} \quad (5)$$

信頼度はビートベクタの自己相関で計算される. 節 3·1 の仮定より, ビートパターンは周期的であるので, ビートトラッキングが正しく行われた時は, スペクトログラムが最新のものと一小節前のものとで同じであるとみなせる. そこで, 信頼度としてビートベクタの一小節長時間ずれの正規化相互相関を導入する. ビートトラッキング手法 X の時刻 t における信頼度 $S_{X,t}$ を以下に示す.

$$S_{X,t} = \frac{\sum_{j=1}^{N_F}\sum_{i=0}^{m_q}\sum_{k=n}^{n-3} q_X(k,i,j)q_X(k-4,i,j)}{\sqrt{\sum_{j=1}^{N_F}\sum_{i=0}^{m_q}\sum_{k=n}^{n-3} q_X(k,i,j)^2 \sum_{j=1}^{N_F}\sum_{i=0}^{m_q}\sum_{k=n-4}^{n-7} q_X(k,i,j)^2}} \quad (6)$$

ただし, n は時刻 t における拍の検出個数であるとする. N_F はビートベクタの次元数であり、本稿では 15 次元すべてを用いた.

以降で村田らや糸原らのビートトラッキング手法で用いられる STPM(Spectro-Temporal Pattern Matching) との比較を行なう. STPM は正規化相互相関関数をオンセットベクタの自己相関に適応したものである. 彼らは自己相関の窓長を減らすことで、短時間のテンポの変動に対応している. 我々の相関関数の時間幅は一小節長である. しかし. ビートベクタにより検出拍による量子化が行なわれているので, テンポ変動の有無を各ビートトラッキング手法に委ねた上で, 短時間のテンポ変動を加味したビートトラッキング結果の選出が可能になる.

本結合手法と類似する例として, 後藤 [10] や Dixon ら [11] の開発したマルチエージェントビートトラッキング手法があげられる. これらのエージェントは各エージェント間で独立にテンポや拍時刻の推定を行なう. Dixon らのビートトラッキング手法ではそれぞれのエージェントが同じ戦略でトラッキングを行うため多様な楽曲に対応できない. また, 後藤のビートトラッキングではそれぞれのエージェントは全音符長, 二分音符長など異なった観点でトラッキングを行うが, その中でオンビート音楽であるという強い仮定を持っている欠点がある. 一方で我々の結合手法では, 各々のエージェントに当たるビートトラッキング手法はそれぞれ独自の仮定を持つことができるので, 様々な楽曲に対応する事が可能になる.

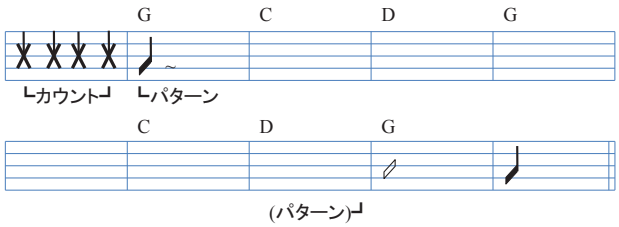

図 3 実験で用いた楽譜の概要図. 中央のパターン部分は図 2 の 8 つのうちのいずれかのパターンが適応される. X, 白音符, 黒音符はそれぞれ, カウントの音, 全音符, 4 分音符を表す.

5. 実験および考察

本結合手法と結合に用いた村田らの手法 [6] と糸原らの手法 [7] の推定精度の比較を行う. この章ではまず実験条件を述べた後に, 使用した村田ら, 糸原らのそれぞれのビートトラッキング手法について述べ, 最後に比較結果を示し, 考察を加える.

5·1 実験条件

ヒューマノイドである HRP-2 を用いて実験を行った. 入力信号はロボットのファンノイズとの混合信号になり, ギター音とノイズは信号対ノイズ比で 5.68dB である. ギター演奏の録音データは, 被験者 4 名でそれぞれテンポ 3 種 (BPM70, 90, 110), ビートパターンは図 2 に示された 8 種である. 順番は, 数字が小さいほど表拍アクセントが, 大きくなるほど裏拍アクセントが多くなるよう設計した. 演奏は図 3 で示されるように 4 つのカウント, 7 回のビートパターンの繰り返し, 最後の全音符音と短音という構成であるとする. カメラの fps は約 19 である. 人とロボットの距離は約 3[m] で, ギター全体が画面に含まれる. また, 推定がビート位置誤差が $\pm 150[msec]$ 以内であるときを推定成功とし, それらの適合率, 再現率をそれぞれ $(r_{prec} = N_e/N_d)$, $(r_{recall} = N_e/N_c)$ で定義する. ただし, N_e, N_d, N_c はそれぞれ推定拍数, 推定成功拍数, 正解拍数を表す. ここで, それらの調和平均である, F 値を導入する:

$$F\text{-}measure = \frac{2}{1/r_{prec} + 1/r_{recall}}. \quad (7)$$

5·2 使用したビートトラッキング手法

5·2.1 村田らのビートトラッキング手法

村田らの手法の構成は, テンポ推定の Spectro-Temporal Pattern Matching(STPM) と拍位置推定のルール適用部である. STPM は STPM は正規化相互相関関数をオンセットベクタの自己相関に適応したもので, 相関の高い自己相関の時間ずれのパラメータが推定テンポとなる. 拍探索ルールでは, 村田らはオンビート音楽を仮定することで, オンセットの強さかつ推定テンポを用いたオンセットの連続性を用いて拍位置を推定している. ここで同様に 8 分音符や 3 連符を除去するための経験則的ルールも用いられている.

表 1 各手法ごとの F 値 (%) の比較 (番号は楽譜パターン)

	1	2	3	4	5	6	7	8	Ave.
結合	0.95	0.90	0.89	0.72	0.63	0.53	0.42	0.33	0.67
村田	0.94	0.90	0.89	0.66	0.64	0.52	0.33	0.31	0.65
糸原	0.81	0.82	0.86	0.78	0.58	0.54	0.84	0.53	0.72

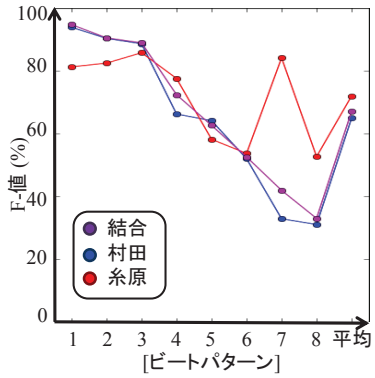

図 4 各手法ごとの F 値 (%) の比較.

村田らの手法の有利な点は, オンセットベクタによるノイズ頑健性, 自己相関の窓幅の小ささによるテンポ変動への頑健性, 仮定であるオンビート音楽での追従性能である. その一方で, オフビート音楽での追従性能が悪さが, 文献 [7] によって明らかになっている.

5·2.2 糸原らのビートトラッキング手法

糸原らの手法はギター演奏者に焦点を置いた視聴覚統合ビートトラッキングである. その構成は, オンセットベクタ及び STPM による音響特徴量, 手とギターの画像トラッキングによる手のギターからの相対距離である画像特徴量, それらの特徴量を統合するパーティクルフィルタ部分になる. 手のギターからの相対距離の特徴は, ビートパターンによらず小節内のどの位置を演奏しているかとの相関性が強いことである. この事実から手の軌道をモデル化し, パーティクルフィルタにより視聴覚統合を行なう. 視聴覚の両方を用いることで, ノイズやテンポ変動性と同時に, 村田らの仮定であるオンビート音楽の制限をのいビートトラッキングを獲得した. その一方で, オンビート音楽では村田らの手法の方が高い精度を出しており, オフビート音楽でもその精度はロボットとの合奏において十分であるとはいえない.

5·3 精度比較

各手法の推定精度の比較結果を表 1 と図 4 に示す. 村田らの手法は, インデックスの小さいオンビートのビートパターンでは高い精度を示す一方, インデックスの増加, つまり裏拍ビートの増加に伴って精度が極端に下がっている. 糸原らの手法では, 手の軌道を 8 ビートでモデル化しているため, それに合致するパターン 1,2,3,4,7 では 80%に近い精度が得られている. しかし, パターン 5,6,8 は 16 ビートであるため推定精度は低く, 80%という数字も決して高いとは言えない.

本結合手法はオンビート音楽及びオンビートとオフビートの混合音楽では比較的よい結果を出している. 例えば, オンビートであるパターン 1 から 3 においては最も高い推定精度を出しており, オンビートとオフビートの混合であるパターン 4 から 6 でも村田らや糸原らの手法の平均のような精度を得ることが確認できた.

その一方で, すべてのアクセント拍がオフビートであるパターン 7,8 では村田らより少し良い程度の結果しか得られなかった. この原因は, 村田らの手法がオフビート音楽において表裏を混同した状態で安定したトラッキングを行なうことである. 現在の信頼度は, ビートトラッキングのビートパターンの周期性のみから計算している. そのため半拍ずれた状態で毎回トラッキングをする村田らの手法でも信頼度上では高い値を得ることができてしまう問題がある.

6. おわりに

本稿では, 音楽ロボットのための人のビートトラッキングの結合手法を報告した. 現在ではまだ大きな改善は見られなかったが, 複数のビートトラッキング手法を結合することで複数の手法の中から逐次的に適切な手法を選ぶことができる可能性が示唆された.

今後, オフビート音楽に対応するために, 検出した拍が表拍か裏拍かを明示的に精査する, 表拍信頼度, 裏拍信頼度の導入の必要が考えられる. また, 実際のロボット合奏による評価も行ないたい.

謝辞 本研究の一部は科研費 (S), の支援を受けた. また, STPM の使用許可をいただいた HRI-JP に感謝します.

参考文献

[1] J. Solis et al. Understanding the mechanisms of the human motor control by imitating flute playing with the Waseda Flutist Robot WF-4RIV. *Mechanism and Machine Theory*, 44(3):527–540, 2008.

[2] T. Mizumoto *et al.* Human-robot ensemble between robot thereminist and human percussionist using coupled oscillator model. In *Proc. of IROS*, pages 1957–1963. IEEE, 2010.

[3] G. Weinberg *et al.* The Creation of a Multi-Human, Multi-Robot Interactive Jam Session. In *Musical Expression*, pages 70–73, 2009.

[4] T. Mizumoto *et al.* Integration of flutist gesture recognition and beat tracking for human-robot ensemble. In *Proc. of IEEE/RSJ-2010 Workshop on Robots and Musical Expression*, pages 159–171, 2010.

[5] G. Hoffman and G. Weinberg. Gesture-based human-robot jazz improvisation. In *Proc. of ICRA*, pages 582–587. IEEE, 2010.

[6] K. Murata *et al.* A beat-tracking robot for human-robot interaction and its evaluation. In *Proc. of Humanoids*, pages 79–84. IEEE, 2008.

[7] T. Itohara *et al.* A multi-modal tempo and beat tracking system based on audio-visual information from live guitar performances. *EURASIP J. on Audio, Speech, and Music Processing*, 2012(1):6, 2012.

[8] Y. Pan *et al.* A robot musician interacting with a human partner through initiative exchange. In *Proc. of the 2010 Conf on New Interfaces for Musical Expression*, pages 166–169, 2010.

[9] A. Lim *et al.* Robot musical accompaniment: integrating audio and visual cues for real-time synchronization with a human flutist. In *Proc. of IROS*, pages 1964–1969, 2010.

[10] M. Goto. An audio-based real-time beat tracking system for music with or without drum-sounds. *J. of New Music Research*, pages 159–171, 2001.

[11] S. Dixon and E. Cambouropoulos. Beat tracking with musical knowledge. In *Proc. of Conf. on Artificial Intelligence*, pages 626–630, 2000.