Vol.2011-MUS-91 No.11
2011/7/28

CRF に基づく伴奏の演奏表現の予測モデルと協調演奏システム

山本龍一[†1]　酒向慎司[†1]　北村正[†1]

本稿では，複数パートを含む楽譜が与えられた際に，演奏者の一部演奏に合わせて，伴奏に適切な演奏表情を付与するための伴奏の予測モデルを提案する．複数パートを含む演奏の場合，それぞれのパートは旋律としての自然さを保ちながら，パート同士が調和して進行すると考える．本研究では，それらの関係を CRF（Conditional Random Fields，条件付き確率場）を用いて統計的に学習し，伴奏の演奏表現の予測モデルの予測精度について評価実験及び考察を行った．また，その応用として実時間で演奏位置を推定し，演奏表情が付与された伴奏を自動再生する協調演奏システムを提案する．

Cooperative Automatic Accompaniment System Using Predictive Models of Expression in Music Performance Based on CRFs

Ryuichi Yamamoto,[†1] Shinji Sako[†1]
and Tadashi Kitamura[†1]

In this paper, we propose the method to predict expression of accompaniment given a part of performance referred to the score which contains several parts. In corroborative performance, a part of performance has musical harmony as a melody and it harmonizes with other parts. In our approach, the harmonic relation between parts is modeled by CRF (Conditional Random Fields). Experimental results and evaluations of the accuracy of the predictive models which learned from data statistically are reported. Also, we present the cooperative automatic accompaniment system which estimates the performer's beat position in the score on real-time processing and play the expressive accompaniment automatically.

1. はじめに

自動伴奏とは，人間の演奏に合わせてコンピュータで伴奏を自動再生させる技術のことを言う．自動伴奏に関する研究は古くから行われており，楽器演奏の技術が未熟な人であっても手軽に演奏を楽しめるようなアンサンブルシステムとして応用されている．また，合唱曲のような複数パートを含む楽曲の練習支援としても寄与する技術であり，現在も盛んに研究がなされている．

本研究では，自動伴奏を人間とコンピュータの協調演奏と捉え，より人間同士の演奏に近い協調演奏を実現することを目標とする．そのような協調演奏を実現するため，厳密に楽譜通りでない演奏者の演奏から楽譜中の演奏位置を求める問題と，演奏者の演奏に合った伴奏の演奏表情を予測する問題の解決に取り組む．

自動伴奏を実現するためには，演奏者の楽譜上の位置をコンピュータで認識する技術が必要不可欠である．しかし，人間の演奏は厳密に楽譜通りに演奏されるとは限らず，局所的なテンポ変化や演奏誤りなどを含むことが多いため，一般には容易ではない．このような問題に対し，隠れマルコフモデル（Hidden Markov Model，HMM）やグラフィカルモデルなどの確率モデルに基づくアプローチが提案されており，演奏ミスを含む演奏に対しても頑健に追従可能なことが報告されている．

人間の演奏は，たとえ同じ楽曲であっても演奏者が違えば異なるものとなり，また同一演奏者でも毎回同じ演奏をするとは限らない．したがって，それに対応する伴奏をどのように再生させるかは一意には決めることができない．しかし，複数パートを含む楽曲の場合は，それぞれのパートはそれ自体で音楽的な自然さを保ちつつ，かつパート毎が調和して一つの楽曲が成り立っていると考えれば，一方のパートから別のパートを予測できる可能性がある．我々は，それらの関係を確率モデルによって捉える．確率モデルを用いることにより，モデルパラメータを統計的に学習可能であり，また入力として想定していないような未知パターンにおいても柔軟に対応可能である．

本稿では，MIDI 演奏を対象とし，柔軟な素性設計が可能な CRF（Conditional Random Field, 条件付き確率場）を用いて伴奏の演奏表現の予測モデルを構築し，その予測精度について評価を行う．また，自動伴奏を実現するにおいて不可欠な楽譜追跡と演奏表現の予測を

†1 名古屋工業大学大学院工学研究科
Graduate School of Engineering, Nagoya Institute of Technology

© 2011 Information Processing Society of Japan

Vol.2011-MUS-91 No.11
2011/7/28

統合した協調演奏システムを提案する．

2. 関 連 研 究

演奏者の演奏を追跡し伴奏を自動再生する自動伴奏は，1984 年頃の Dannenberg と Vercoe らの研究[1),2)] に始まり，以降急速に発展している．ここでは，本研究に取り分けて関係のあると思われる研究を簡単にまとめる．

武田らは，多重音 MIDI 演奏において，演奏者の演奏を HMM でモデル化し，局所的なテンポ変化・演奏誤り・弾き直しに対応可能な自動伴奏を提案した[3)]．また，鈴木らが単旋律の音響入力に拡張し，スペクトル形状の時間変化が少ない楽器に対して音響入力で対応可能であると報告している[4)]．Raphael らは，入力演奏と伴奏を共に音響演奏を対象とし，演奏生成の HMM と伴奏のタイミングの予測モデルを用いてテンポ変化や演奏ミスに頑健な自動伴奏を提案している[5)]．これらの研究は，演奏ミスを含む人間の演奏に伴奏がどのようにして追従すべきかに重点を置いており，伴奏の演奏表現をどうするべきかといった問題については十分に議論されておらず，協調演奏を指向したものではない．

一方，堀内らは，人間同士の合奏を分析し，重回帰分析による伴奏の演奏表現の予測を用いた自動伴奏システムの構築を行った[6)]．また，主観評価によりシステムの性能評価を行った[7)]．

本研究では，人間とコンピュータの協調演奏を指針として，確率モデルに基づく伴奏の演奏表現の予測手法に重点を置いて議論する．

3. HMM に基づく MIDI 演奏の楽譜追跡

3.1 楽 譜 追 跡

演奏者がある楽譜に従って楽曲を演奏する場合に，楽譜中のどの位置を演奏しているかを求める問題を楽譜追跡と呼ぶ．人間が楽譜通りに正しく演奏することが保証されていれば，入力音と楽譜上の音を時系列に従って対応を取ることにより，楽譜追跡は容易に実現できる．しかし，実際の人間の演奏には演奏ミスやテンポ変動などの楽譜との不一致を含むため，単純な時系列マッチングではうまくいかないことが多い．本研究では，鍵盤楽器を用いた MIDI 演奏を対象とし，武田らによって提案されている演奏生成の確率モデルに基づく柔軟な楽譜追跡手法について述べる．

3.2 HMM による演奏のモデル化

電子ピアノなどを用いた MIDI 演奏では，ハードウェアの遅延はほぼ無視できるとすると，音の高さと音の長さに相当する情報が瞬時に取得できる．和音を一つのクラスタとし，楽譜上の音符が指定されている位置を s_i とする．また，s_i において演奏者がある音符 n_i を一つ前の拍から t_i 秒後に発音したとし，$x_i = (n_i, t_i)$ とおく．t_i は 前の拍から当該拍へと遷移する際の発音時間間隔であり，IOI（Inter Onset Interval）と呼ぶ．

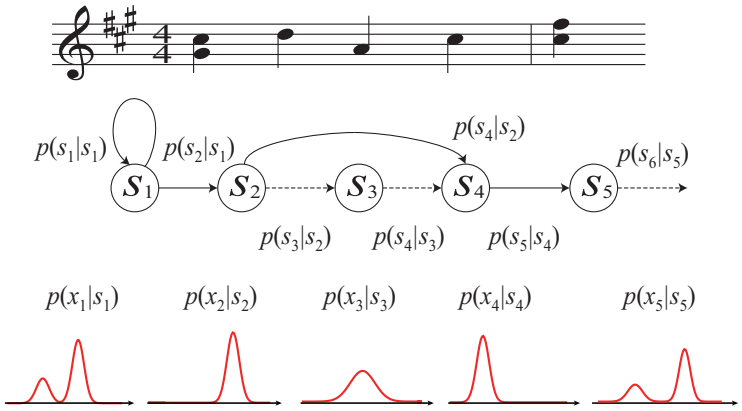


図 1 演奏生成の隠れマルコフモデル
Fig. 1 Hidden markov models which represent human performance

観測系列を $\boldsymbol{X} = \{x_1, x_2, ..., x_n\}$，楽譜上の位置を $\boldsymbol{S} = \{s_1, s_2, ..., s_n\}$ とすると，楽譜に従った演奏というのは，それらを結ぶ**図 3.2** のようなネットワークを遷移しながら演奏を生成する過程とみなすことができる．ここで $p(s_i|s_j)$ は j 番目の拍位置から i 番目の拍位置へと遷移する確率であり，$p(x_i|s_i)$ は i 番目の拍位置で x_i を出力する確率である．人間の演奏は演奏誤り，時には弾き直しなどを含むが，それらの傾向は確率として記述可能である．なお，状態数は楽曲の長さに比例して大きくなり，また演奏の傾向も演奏者・楽曲によって異なるため，現実的には学習は困難となる．しかし，楽譜中で C（ドの音）が指定されている箇所では，C が一番演奏され易く，また隣り合う D（レの音）と B（シの音）が誤って演奏されやすいなど，経験的に確率を与えることができる．和音は理想的には同一に演奏されるべきであるが，実際には小さな IOI の値を取る．トリルなどの装飾音を含まない演奏においては閾値判定で十分に和音かどうかを判断可能である．

3.2.1 事後確率最大化としての定式化

入力される人間の演奏に対して，最も確率が高くなるような楽譜の遷移を求めたい．前節

© 2011 Information Processing Society of Japan

で述べた通り，人間の演奏は楽譜の位置を隠れ状態とした HMM によって生成されると考える．すると，観測系列を $\boldsymbol{X} = \{x_1, x_2, ..., x_n\}$ ，楽譜上の位置を $\boldsymbol{S} = \{s_1, s_2, ..., s_n\}$ を求める問題は，事後確率最大化として式 (3) として定式化できる．

$$\hat{\boldsymbol{S}} = \arg\max_{S} P(\boldsymbol{S}|\boldsymbol{X}) \tag{1}$$

$$= \arg\max_{S} P(\boldsymbol{S}|\boldsymbol{X})P(\boldsymbol{S}) \tag{2}$$

$$= \arg\max_{S} \prod_{i=1}^{n} p(x_i|s_i)p(s_i|s_{i-1}) \tag{3}$$

すべての可能な $\boldsymbol{S}$ に対して事後確率が最大となる拍遷移系列を求める際，拍数が多いほど組み合わせは膨大になるため，単純に求めるのは困難である．そこで，効率よく最尤な拍位置の遷移を求めるために，HMM における状態系列を求める手法として広く使われている Viterbi アルゴリズムを用いる．Viterbi アルゴリズムを用いることによって，計算量は状態数に対して線形オーダーとなり，実時間で最尤拍遷移を求めることが可能となる．

4. CRF に基づく伴奏の予測モデル

4.1 協調演奏

複数パートを含む楽曲に対して演奏者が演奏するパート以外の部分を伴奏と呼ぶ．楽譜を既知とし，人間に合わせて表情を持った伴奏を自動再生する協調演奏への応用を想定し，ここでは，ある楽曲に対して一部パートの演奏が与えられたとき，伴奏をどのように演奏すべきかといった問題を扱う．

一般的に，人間同士の演奏においては両者が互いの音を聴きながら進行していくものである．例として，2 名からなる最小構成のアンサンブルを考える．一方の演奏者のテンポが速くなれば，もう一方の演奏者もテンポを速くし，一方が曲を盛り上げるように演奏すれば，もう一方も盛り上げるように演奏するのが自然であろう．本研究では，伴奏の演奏表現は次の 2 つを満たすように決定されると考える．

(1) 伴奏のみ着目したときに，音楽的に自然である
(2) 演奏者の演奏に調和している

伴奏は，それ自体が自然な旋律として演奏される一方で，演奏者の演奏の影響を受けて演奏される．それらは相互に独立ではなく，同時に扱われるべきであると考える．

音楽演奏は，音符単位の局所的な演奏の系列であるとみなすことができる．ただし，音楽演奏は時系列的な特徴を持っており，ある時刻での演奏はその前後の演奏に依存していると考えられる．例えば，徐々に音階が上がっていく，またはテンポが上がっていくような演奏では，徐々に音量が上がり傾向になることが考えられる．また，音楽にはフレーズという時系列上の複数の音符からなる旋律における自然な区分が一般的に存在し，フレーズ毎にある程度演奏の特徴を持っている．

本研究では，局所的な演奏表現の対応関係モデル化するのに素性関数として柔軟に設計が可能な CRF を用いる．CRF は，入力系列全体が与えられたときの出力系列に対する条件付き確率をモデル化する手法である．演奏者の演奏表現が得られたときに，伴奏の演奏表現を予測する問題を CRF を用いて事後確率最大化問題として定式化する．本報告では，その初歩段階として，演奏表現において重要であると考えられる音量を予測するモデルを獲得する手法について述べる．

4.2 CRF を用いた予測モデル

本研究は，自動伴奏システムへの応用を念頭に置いており，伴奏は MIDI 信号を自動再生することによって演奏される．したがって，演奏表情は MIDI 信号で表現可能なものを仮定する．MIDI 信号における音量は 0 ～ 127 の離散値で与えられる．以降の議論では，入力系列としての演奏者の演奏した音符系列から得られる演奏表情を $\boldsymbol{X} = \{x_1, x_2, ..., .x_n\}$ とし，出力として伴奏の演奏表情を $\boldsymbol{Y} = \{y_1, y_2, ..., .y_n\}$ とする．

伴奏の局所的な音符の演奏表情は，ある時間区間の平均からの逸脱であると考え，伴奏の音量の特徴量として，各音符に対して次のように定義する．ただし，v_i^A は伴奏パートの音符 n_i^A に対する音量を表し．v_{ave}^A は n_i^A の周辺のある区間に含まれる音符の平均音量を表す．

$$y_i = \log \frac{v_i^A}{v_{ave}^A} \tag{4}$$

演奏者の演奏表現の特徴量を考える．音符系列は離散的であり，伴奏の音符が演奏されるべき位置において必ずしも演奏者がある音符を演奏するとは限らない．したがって，伴奏の音符が指定されている時点での演奏者の演奏表現として，ある時間区間における平均音量とその時間微分に相当する次式のような特徴量を用いる．

$$x_i = \frac{1}{N} \sum_{j=1}^{N} v_{i-j}^P \tag{5}$$

$$\Delta_i = x_i - x_{i-1} \tag{6}$$

ただし，v_i^P は演奏者パートの音符 n_i^P に対する音量を表し，N は n_i^P が演奏されるまでのある区間内に含まれる音符数とする．

© 2011 Information Processing Society of Japan

表 1 CRF で使用する素性
Table 1 score features

演奏者	平均音量 $+\Delta$	-
伴奏	平均音量からの逸脱	音高の変化

伴奏の演奏表情系列 $\boldsymbol{Y}$ と演奏者の演奏表情系列 $\boldsymbol{X}$ は，**表 1** に示される楽譜素性の組み合わせによって素性関数 $\phi(\boldsymbol{X}, \boldsymbol{Y})$ が記述される．演奏者の演奏表情に対する伴奏の演奏表情の確信度を表す素性ベクトルは，素性関数の重み付き和として次式のように与えられる．Θ は各素性関数の重み係数の集合であり，CRF のパラメータである．

$$\langle \Theta, \Phi(\boldsymbol{X}, \boldsymbol{Y}) \rangle = \sum_{f \in F} \theta_f \phi_f(\boldsymbol{X}, \boldsymbol{Y}) \tag{7}$$

ここで，演奏表情 y_i が一つ前の y_{i-1} のみに依存すると仮定すると，演奏者の演奏表情系列 $\boldsymbol{X}$ が与えられたときの伴奏の演奏表情系列 $\boldsymbol{Y}$ は Linier Chain CRF として次のようにモデル化できる．

$$P(\boldsymbol{Y}|\boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})} \exp\langle \Theta, \Phi(\boldsymbol{X}, \boldsymbol{Y}) \rangle \tag{8}$$

$$= \frac{1}{Z(\boldsymbol{X})} \exp \sum_{f \in F} \theta_f \phi(\boldsymbol{X}, \boldsymbol{Y}) \tag{9}$$

ただし，$Z(\boldsymbol{X})$ は正規化項であり，次式を満たすとする．

$$Z(\boldsymbol{X}) = \sum_{Y'} \exp\langle \Theta, \Phi(\boldsymbol{X}, \boldsymbol{Y}') \rangle \tag{10}$$

N 個の学習データが与えられたとき，パラメータ Θ は次式のように尤度が最大となるように学習される．学習には，一般的には準ニュートン法や stochastic gradient decent などの効率の良いアルゴリズムが用いられる．

$$\hat{\Theta} = \arg\max_{\Theta} \prod_{i=1}^{N} P(\boldsymbol{Y}_i|\boldsymbol{X}_i; \Theta) \tag{11}$$

予測に関しては，次式のように事後確率が最大となるように求められ，Viterbi アルゴリズムを用いることで高速に計算できる．

$$\hat{\boldsymbol{Y}} = \arg\max_{Y} P(\boldsymbol{Y}|\boldsymbol{X}) \tag{12}$$

表 2 CRF の学習に用いた楽曲
Table 2 training samples

楽曲名	演奏者	合計
Piano Sonata K.331 1st Mov.	C.Eschenbach, G.Gould, I.Haebler, L.Kraus, N.Shimizu, H.Nakamura, A.De Larrocha, M.J.Pires	8 曲

5. 評価実験

協調演奏システムでは，多様なユーザの演奏に対応するため，演奏者の違いによってシステムが柔軟に対応可能であることが望まれる．そこで，提案手法の有効性を確認するために，演奏者の違いによって予測精度がどの程度保たれるのかを検証する実験を行った．学習に用いる実演奏データベースには CrestMusePEDB[⋆1] の v2.2 及び v2.3 を用いた．なお，CRF の学習には stochastic gradient decent アルゴリズム[8] の実装である crfsgd[⋆2]を用いた．

5.1 実験条件

表 2 に示すように W. A. Mozart 作曲の Piano Sonata K.331 1st Mov. に対して，複数の演奏者が演奏した楽曲を学習に用いた．それぞれの楽曲を右手と左手のパートに分け，左手のパートを伴奏とした．実験では，右手の演奏から左手の演奏を予測する．なお，和音を一つのクラスタとし，和音に含まれる音符はすべて同一の演奏表情を持つして扱う．評価データとしては，**表 3** に示す 3 セットを用いる．提案手法が演奏者の違いに頑健であるかどうかを検証するため，評価データには異なる演奏者の楽曲を用いた．また，学習データに含まれない楽曲も評価データとして用い，未知の楽曲に対して提案手法がどのように働くかを検証する．推定する伴奏の音量を表すパラメータは，k-means アルゴリズムを用いて 16 段階に量子化した．k-means アルゴリズムを用いることにより，パラメータ空間をより良く近似するように非線形量子化が可能である．

5.2 同一楽曲で異なる演奏者の場合の実験結果

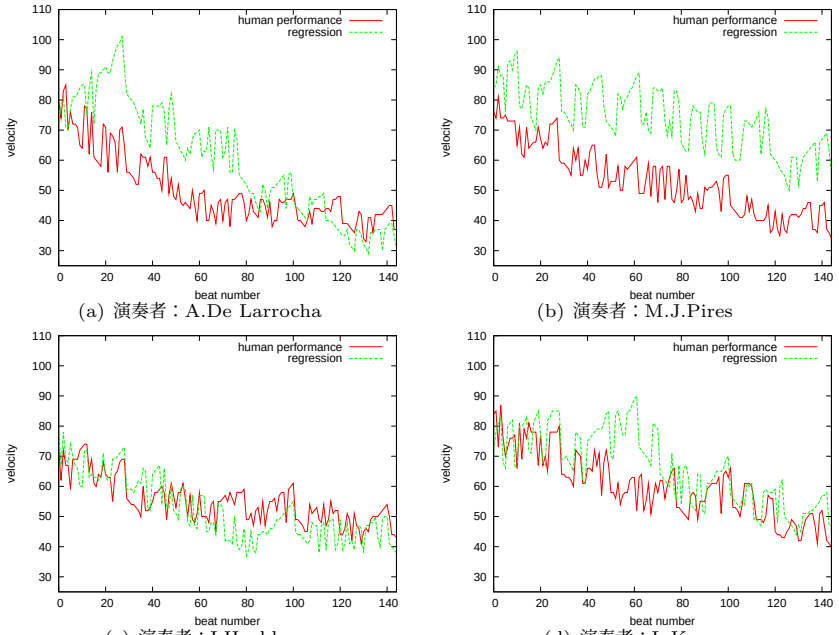
図 2 の中から評価データを除いた演奏者 6 名分の演奏データでモデルを学習し，残りのデータで予測実験を行った．**図 2** に，学習に用いた楽曲と同一の楽曲で，異なる演奏者の演奏に対して伴奏の音量の予測を行った結果を示す．横軸は伴奏の拍数であり，縦軸は音量（0〜127）である．

⋆1 CrestMusePEDB: http://www.crestmuse.jp/pedb
⋆2 crfsgd: http://leon.bottou.org/projects/sgd

© 2011 Information Processing Society of Japan

表 3 評価に用いた楽曲
Table 3 test samples

楽曲名	演奏者	合計
Wohltemperierte Klavier I-1 BWV846 Prelude	G.Gould, F.Gulda	2 曲
Piano Sonata K.331 1st Mov.	A.De Larrocha, M.J.Pires	2 曲
Piano Sonata K.331 1st Mov.	I.Haebler, L.Kraus	2 曲

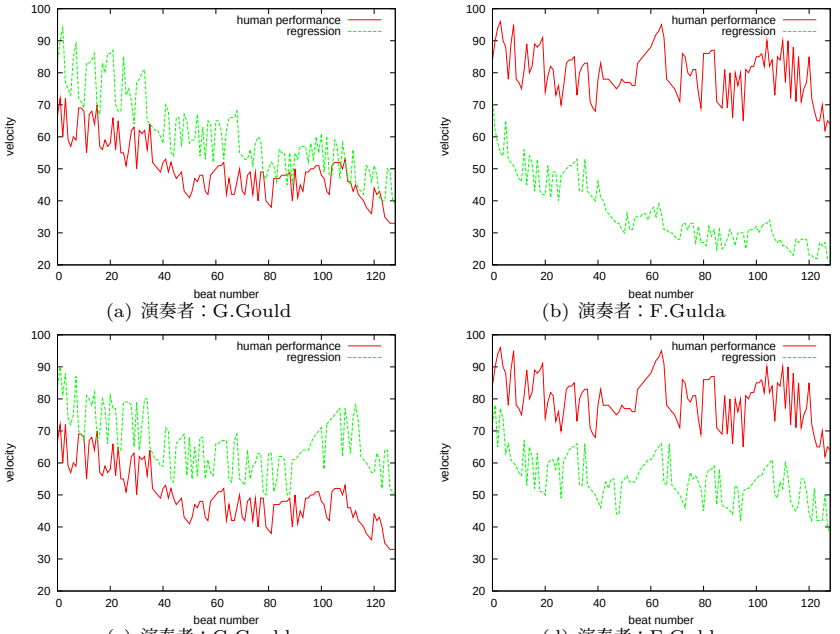


(a) 演奏者：A.De Larrocha (b) 演奏者：M.J.Pires
(c) 演奏者：I.Haebler (d) 演奏者：L.Kraus

図 2 学習データと同一の楽曲に対する伴奏の音量の予測結果：Piano Sonata K.331 1st Mov.
Fig. 2 Experimental results of the regression of velocoity for test samples which are same song as training data: Piano Sonata K.331 1st Mov.

5.3 楽曲が未知の場合の実験結果

前節と同様に，演奏者 6 名分の演奏データからモデルを学習し，未知楽曲である J. S. Bach 作曲の Wohltemperierte Klavier I-1 BWV846 Prelude に対して予測実験を行った．**図 3** に，学習に用いたものとは異なる楽曲を用いて，伴奏の音量の予測を行った結果を示す．



(a) 演奏者：G.Gould (b) 演奏者：F.Gulda
(c) 演奏者：G.Gould (d) 演奏者：F.Gulda

図 3 学習データとは異なる楽曲に対する伴奏の音量の予測結果
：Wohltemperierte Klavier I-1 BWV846 Prelude
Fig. 3 Experimantal rresults of the regression of velocoity for test samples NOT included in trainng data: Wohltemperierte Klavier I-1 BWV846 Prelude

5.4 考　　察

図 2 を見ると，演奏者によって違いはあるが，大まかな音量の変動は捉えられていることがわかる．学習に用いた楽曲と同一の評価データに対しては，演奏者が異なってもある程度予測できることがわかった．一方で，実際の演奏の音量とは大きくかけ離れている部分も見られた．今回は，伴奏の音量の特徴量としてある区間における平均からの逸脱量としたため，一度ずれるとそれ以降にずれが波及してしまうことが原因だと考えられる．

図 3 を見ると，学習データに含まれない楽曲であるにもかかわらず，音量の変動の幅は大まかに予測できている．しかし，同一楽曲の場合と同様に，一度生じたずれがそれ以降に波及していると思われる箇所が多く見られ，また学習に用いる楽曲に応じて精度も大きく変化

© 2011 Information Processing Society of Japan

した．このことから，未知楽曲に対しては，別のパートと調和するといった部分が予測できていない可能性があるが，旋律における音楽的な自然さはある程度予測できていると考えられる．

以上の結果より，提案したモデルでは，学習に用いた楽曲と同一の評価データに対しては演奏者の違いに大まかに対応可能であり，また学習データに含まれない未知楽曲に対してもある程度対応可能であることが示唆された．

6. おわりに

本研究では，複数パートを含む楽曲の演奏において，各パートはそれ自体が自然な旋律として演奏される一方で，パート同士が調和するように進行すると考え，演奏者の演奏から伴奏の演奏表現を予測する予測モデルの構築を行った．演奏表現として重要と考えられる音量において予測実験を行い，モデルの妥当性を検証した．実験結果より，学習データと同一の楽曲であれば演奏者の違いにも対応可能なモデルであることが示唆され，学習データに含まれない未知楽曲であっても対応できる可能性があることがわかった．

本研究において，一部パートが演奏されたときに，その演奏から伴奏の演奏表現を予測し自動再生する協調演奏システムを提案した．入力とする人間の演奏に対して，データから学習した予測モデルに基づいて適切な演奏表情を持った伴奏を自動再生することにより，より人間同士の演奏に近いものを実現可能である．

今後の課題としては，提案モデルの有効性を示すために，予測モデルの精度に関して回帰モデルなど他のモデルと比較実験を行うことを考えている．また，演奏表情は非常に多様なため，精度による評価が必ずしも適切であるとは言えない．したがって，人間の聴取による主観評価を行い，提案する予測モデル及び協調演奏システムの性能を評価する必要があると考えている．予測モデルについては，音量以外にもテンポ，音長の変動に対応するモデルを構築する必要があるほか，協調演奏の予測に有効と考えられる素性に関して十分に検討を行っていきたい．

謝辞 本研究の一部は，文部科学省科学研究費補助金（課題番号: 21700191) ならびに名古屋工業大学平成 23 年度学内研究推進経費の支援を受けて行われたものである．

参考文献

1) Dannenberg, R.B.: An On-line Algorithm for Real-time Accompaniment, *Proc. of the International Computer Music Conference*, pp.193–198, (1984).
2) B. Vercoe.: The Synthetic Performer n the Context of Live Performance, *Proc. of the International Computer Music Conference*, pp.199–200,（1984）.
3) 武田 晴人，西本 卓也，嵯峨山 茂樹.: HMM による MIDI 演奏の楽譜追跡と自動伴奏，情報処理学会研究報告（MUS），pp109–116,（2004）.
4) 鈴木 孝輔，上田 雄，斎藤 康之，小野 順貴，嵯峨山 茂樹.: HMM を用いた音響演奏の楽譜追跡による引き直しに追従可能な自動伴奏，情報処理学会研究報告（MUS），pp1–6,（2011）.
5) Christopher Raphael.: Orchestra in a Box: A System for Real-Time Musical Accompaniment, *Proc. of the International Joint Conferences on Artificial Intelligence* , (2003).
6) 堀内 靖雄，坂本 圭司，市川 熹.: 合奏時の人間の演奏制御の分析・推定，情報処理学会論文誌，vol.45，No.3，pp690–697,（2004）.
7) 矢島 直人，堀内 靖雄，西田 昌史，市川 熹.: 人間の伴奏制御モデルに基づく伴奏システムの実装と評価，情報処理学会研究報告（MUS），pp37-42,（2005）.
8) Bottou, L.: Stochastic Gradient Learning in Neural Networks, *Proc. of the Neuro-Nîmes 91*, Nimes, France, EC2（1991）.

© 2011 Information Processing Society of Japan