Vol.2018-MUS-120 No.19
2018/8/23

多重音検出とリズム量子化の統合による多声音楽の自動採譜

中村 栄太[1,a)]　Emmanouil Benetos[2]　吉井 和佳[1]　Simon Dixon[2]

概要：自動採譜の多くの研究では，出力はピアノロール形式であり，音楽的に解釈されたリズムや音高を表していない．本研究では多重音検出とリズム量子化手法を統合して，多声音楽音響信号を人間が読める楽譜に変換する自動採譜手法を論じる．この統合においては，多重音検出の結果は余分な音符や時間ずれなどの誤りを含むことが問題になる．そこで本研究では拍節隠れマルコフモデルを拡張して，余分な音符を除去できるリズム量子化手法を提案する．また，反復音の取り扱いと発音時刻の調節をするために多重音検出手法の改良を行う．さらに，自動採譜結果を評価するための評価手法を提案する．クラシックピアノ音楽データでを用いた評価により，これらの取り扱いが採譜性能が向上に有効であることを示す．

1. はじめに

音楽音響信号から楽譜への変換である自動音楽採譜は音楽情報処理の根本的な課題の一つである [1, 2]．楽譜の音符は半音単位で離散化された音高と拍単位で離散化された発音および消音時刻（「発音楽譜時刻」および「消音楽譜時刻」）によって表されるため，これらの情報を音響信号から認識することが必要となる．統計音声認識 [3] とのアナロジーによる一つの方法は，楽譜モデルと音響モデルの統合によるものである [4]．しかし，同時音高の組み合わせ数が膨大である問題により，多声音楽に対してはこの方法は現状では実現が難しい．そこで，音高の離散化である多重音検出と発音・消音楽譜時刻を認識するリズム量子化を別々に行う手法がこれまで多く研究されている．

多重音検出手法は多声音楽音響信号を入力として，秒単位の発音および消音時刻と音高，ベロシティー（音強）で表される音符列（これを「音符トラックデータ」と呼ぶ）を出力することで，各時間フレームにおける音高の配置を記述するものである．現在の最高性能の手法は，スペクトログラム分解と深層学習に基づく手法の大きく 2 種類に分類される．スペクトログラム分解法は通常，入力スペクトログラムを各音高などに対応するスペクトルテンプレートを表す「基底行列」と時刻ごとの音高成分の強度を表す「アクティベーション行列」に分解する．代表的手法として，NMF（non-negeative matrix factorization; 非負値行列分解）や PLCA（probabilistic latent component analysis; 確率潜在要素分析）やスパース符号化などがある [5–7]．多重音検出に深層学習を用いた手法として，これまで feed-forward 型や再帰型や畳み込みネットワークが応用されてきた [8, 9].

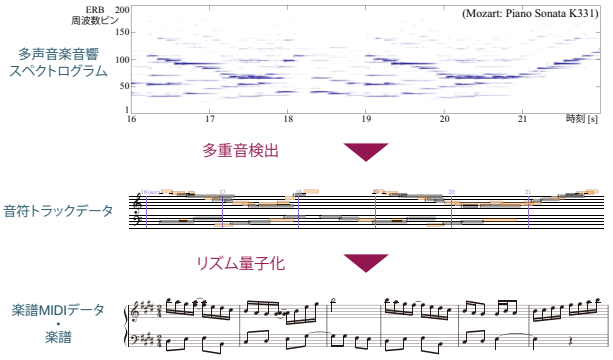

図 1　多重音検出とリズム量子化の統合による多声音楽採譜手法

リズム量子化手法は音符トラックデータまたは人間の演奏 MIDI データを入力として，拍単位で離散化された発音・消音楽譜時刻を持つ音符からなる「楽譜 MIDI データ」を出力する．発音楽譜時刻の推定は通常入力データに含まれる時間変動を取り除くことによって行われる．これまで人手による規則に基づく手法 [10, 11] や統計モデルに基づく手法 [12–18] などが研究されている．最近の研究 [18] では，HMM（hidden Markov model; 隠れマルコフモデル）を用いた手法が現在の最高性能を持つことが示された．特に，拍節 HMM [13, 14] を用いた手法は，拍子と小節線の推定ができる点と不完全な三連符などの音楽文法上間違った楽譜表示を避けられる点で好都合である．消音楽譜時刻（あるいは音価）の認識に関しては，マルコフ確率場を用いた手法が現在最高性能を達成している [19].

[1] 京都大学
Kyoto 606-8501, Japan
[2] Queen Mary University of London
London E1 4NS, UK
[a)] enakamura@sap.ist.i.kyoto-u.ac.jp

多重音検出とリズム量子化手法の最近の発展を踏まえ，本稿ではこれらの統合による音響信号から楽譜への変換が可能な多声音楽採譜手法について述べる（図 1）．これに向けて，時間フレームベースの多重音検出部分を改良し，後段のリズム量子化により適した，音楽的に意味のある結果を出力できるようにする．音符トラックデータは通常，余分な音符（いわゆる false positives）などの正解楽譜には含まれない誤り音符を含んでいるため，誤りの蓄積を避けるためには，リズム量子化手法においてこれらの誤り音符を提言する必要がある [20]．もう一つの課題は，音楽演奏表情に基づく時間ずれに加えて多重音検出で加わる時間誤りを含む音符トラックデータに対してリズム量子化手法のパラメータを適応することである．さらに，自動採譜のプロセス全体に対する評価の方法論を開発することも必要である（ [21] も参照）．

本研究の貢献は以下の通りである．第一に，ここで開発する音響信号から離散記号からなる楽譜への変換を行う多声音楽採譜システムは，我々が知る限り文献の中で初めての試みである．第二に，音符トラックデータに含まれる余分な音符を低減する新規のリズム量子化手法を提案する．時間の規則構造など楽譜の音符に関するトップダウン的知識を取り入れるため，楽譜由来の音符を記述する拍節 HMM と余分な音符の生成を表すノイズモデルの混合過程を表す「Noisy 拍節 HMM」と呼ぶ生成モデルを構築する．第三に，リズム量子化手法のパラメータを最適化して，その効果を検証する．第四に，発音時刻の調整と反復音の検出に関して PLCA に基づく多重音検出手法 [7] を改良を行う．第五に，正解楽譜に基づいて推定楽譜を評価する方法を提案し，一般的に用いられるクラシックピアノ音楽データを用いた系統評価の結果を示す．上記の改良が採譜精度の向上に有効であることを確認し，また提案のリズム量子化手法最が公開されているソフト（MuseScore 2 [23] および Finale 2014 [24]）に比べ有意に性能が優っていることを示す．

2. システム構成

図 2 に提案の多声音楽採譜システムの構成を示す．システム構成は一般の多声音楽にも適用可能なものであるが，一部の要素はピアノ採譜のために適応させている．システムは多重音検出とリズム量子化の 2 つの主な要素からなる（1 節も参照）．

多重音検出部分（3 節）は，各時間フレームごとに多重音のアクティベーションを推定する「多重音解析」と発音・消音時刻，音高，ベロシティーで指定される音符を検出する「音符トラッキング」からなり，出力は音符トラックデータである．リズム量子化部分は発音楽譜時刻を推定する「発音時刻リズム量子化」（4 節）と消音楽譜時刻を推定する「音価認識」からなる．音価認識には，マルコフ確率場に基づく手法 [19] を用いる．また，楽譜 MIDI データに楽譜の上下段に相当する両手パートの情報を含ませるため，文献 [25] の両手パート分離手法を用いる．

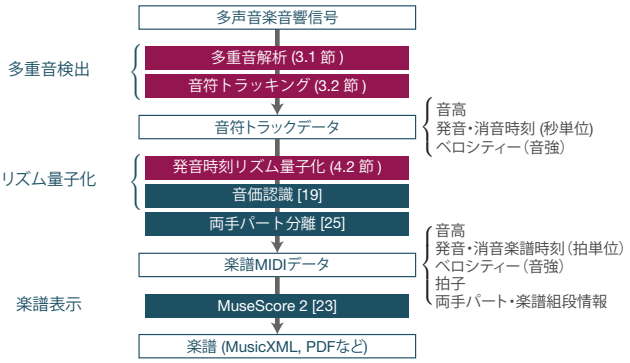

図 2　提案システムの構成

最後に，MusicXML や PDF 形式での楽譜形式を得るため，楽譜編集ソフトの MIDI インポート機能を適用する．具体的には，楽譜の各段の中で声部分離を自動で行う機能を持つ MuseScore 2 [23] を用いる．

3. 多重音検出

3.1 多重音解析

本研究の音響モデルは，スペクトログラム分解にを用いた文献 [7] の手法に基づく．この手法は PLCA [26] の拡張モデルに基づき，入力として ERB（equivalent rectangular bandwidth）スペクトログラムを用いる．周波数のインデックスを ω，時間フレームのインデックスを t として，ERB スペクトログラムを $V_{\omega,t}$ で表すことにする．スペクトログラムのフィルター数は $\Omega = 250$ であり，周波数ビンは ERB スケール上で 5 Hz and 10.8 kHz の間で線形に等間隔に配置し，ホップサイズは 23 ms としたものを用いる．本研究で，文献 [7] で使われた VQT（variable-Q transform; 可変 Q 変換）スペクトログラムではなく ERB スペクトログラムを用いる理由は，後者は時間分解能に優れたよりコンパクトな表現が可能なことによる．

音響モデルは，入力 ERB スペクトログラムを 2 変数の確率分布 $P(\omega, t)$ で近似する．この確率を，以下のように音高，楽器音源，音状態に依存するアクティベーションの確率の積として分解することを考える．

$$P(\omega,t) = P(t)\sum_{q,p,i} P(\omega|q,p,i)P_t(i|p)P_t(p)P_t(q|p). \quad (1)$$

ここで，p は音高を表し（$p \in \{1 = \mathrm{A0}, \ldots, 88 = \mathrm{C8}\}$），$q \in \{1, \ldots, Q\}$ は音状態（本研究ではアタック，サステイン，リリース状態に対応する $Q = 3$ 状態を考える），$i \in \{1, \ldots, I\}$ は楽器音源を表す（本研究では $I = 8$ として 8 種類のピアノを考える）．$P(t)$ は $\sum_\omega V_{\omega,t}$ に対応し，既知の量である．$P(\omega|q,p,i)$ は，楽器 i，音高 p，音状態 q ごとに事前学習するスペクトルテンプレートの辞書に対応す

る．$P_t(i|p)$ は，音高 p に対する各時刻での楽器 i の寄与を表し，$P_t(p)$ は音高のアクティベーション，$P_t(q|p)$ は各音高・各時刻における音状態のアクティベーションを表す．

未知のパラメータである $P_t(i|p)$, $P_t(p)$, $P_t(q|p)$ は EM アルゴリズム [27] を用いて逐次推定する．辞書 $P(\omega|q,p,i)$ は，ここでは固定されたものとして扱い，更新はしない．文献 [7] と同様に，採譜結果における同時音高数や楽器音源の寄与を制御するために $P_t(p)$ と $P_t(i|p)$ にスパース制約を設ける．多重音解析の出力は，音高アクティベーション確率と振幅スペクトログラムの積 $P(p,t)=P(t)P_t(p)$ で与えられる．

3.2 音符トラッキング

音符トラッキング部分は，連続値をとる時間音高表現 $P(p,t)$ を発音・消音時刻を持つ音符のリストに変換する．このため，$P(p,t)$ を閾値処理によりバイナリ化して，30 ms 未満の長さを持つ音符を取り除く．この結果に対して，以下のように反復音符の検出を行う．まず，音符検出された時間周波数領域における $V_{\omega,t}$ のピークを検出する（この際，検出された音符の基本周波数に対応する周波数ビンのみを用いる）．検出されたピークは反復音符の発音を示唆するため，これに基づき検出音符を細分化する．最後に，発音時刻の微調整を行うため，検出された音符の発音時刻周辺での $V_{\omega,t}$ のスペクトルフラックス特徴量のピークを検出する．具体的には，検出された発音時刻の前後 50 ms の領域でピーク検出を行い，発音時刻の微調整を行う．

4. 発音時刻リズム量子化

4.1 発音時刻リズム量子化のための拍節 HMM

まず，拍節 HMM [13,14] について解説する．このモデルは楽譜モデルと演奏タイミングモデルからなる．楽譜モデルは各音符の発音楽譜時刻に対応するビート位置（小節内での相対的な発音楽譜時刻）を生成する．音符のインデックスを n ($n=1,\dots,N$)，そのビート位置を $b_n\in\{0,\dots,B-1\}$（B は小節の長さを表す）と表す．和音を表現するために，$(n-1)$ 番目と n 番目の音符が同じ和音に属する（$g_n=\mathrm{CH}$）か否か（$g_n=\mathrm{NC}$）を表す二値変数 g_n（「和音変数」と呼ぶ）を用いる．$b_{1:N}$ と $g_{1:N}$ は，初期確率 $P(b_1,g_1)$ と遷移確率 $P(b_n,g_n|b_{n-1})$ により生成される．この際，$g_n=\mathrm{CH}$ ならば $b_n=b_{n-1}$ となる制約を満たすものとする．$(n-1)$ 番目と n 番目の音符の楽譜時刻の差は以下で与えられる．

$$[b_{n-1},b_n,g_n]=\begin{cases}0, & g_n=\mathrm{CH};\\ b_n-b_{n-1}, & g_n=\mathrm{NC},\ b_n>b_{n-1};\\ b_n-b_{n-1}+B, & g_n=\mathrm{NC},\ b_n\le b_{n-1}.\end{cases}$$

演奏タイミングモデルは，$t_{1:N}$ で表される発音時刻を生成する．テンポ変動を表現するため，以下のガウス・マルコフモデルに従う局所テンポ変数 $v_{1:N}$ を導入する．

$$v_1=\mathrm{Gauss}(v_{\mathrm{ini}},\sigma^2_{\mathrm{ini}\,v}),\quad v_n=\mathrm{Gauss}(v_{n-1},\sigma_v^2).\quad(2)$$

ここで，$\mathrm{Gauss}(\mu,\Sigma)$ は平均 μ，分散 Σ を持つガウス分布を表し，v_{ini} は初期テンポの参照値，$\sigma_{\mathrm{ini}\,v}$ は大局的テンポのゆらぎの大きさを表す標準偏差，σ_v はテンポ変動の大きさを表す標準偏差である．n 番目の音符の発音時刻 t_n は，一つ前の発音時刻 t_{n-1} と変数 v_{n-1}, b_{n-1}, b_n, g_n に依存した以下の確率により生成される [18]．

$$t_n=\begin{cases}\mathrm{Gauss}(t_{n-1}+v_{n-1}[b_{n-1},b_n,g_n],\sigma_t^2), & g_n=\mathrm{NC};\\ \mathrm{Exp}(t_{n-1},\lambda_t), & g_n=\mathrm{CH}.\end{cases}\quad(3)$$

ここで，$\mathrm{Exp}(x,\lambda)$ はスケールパラメータ λ と定義域 $[x,\infty)$ を持つ指数分布を表す．発音時刻リズム量子化は，入力系列 $t_{1:N}$ から変数 $b_{1:N}$, $g_{1:N}$, $v_{1:N}$ を推定することにより行えるが，これにはテンポ変数の離散化の後に Viterbi アルゴリズムを適用できる．

4.2 Noisy 拍節 HMM

Noisy 拍節 HMM は，拍節 HMM とノイズモデルを組み合わせて構成される．ノイズモデルでは，発音時刻は以下のように生成される．

$$P_*(t_n|t')=\mathrm{Gauss}(t_n;t',\sigma_*^2),\quad(4)$$

ここで，σ_* は標準偏差を表し，σ_t よりも大きいと想定される．また参照時刻 t' には，以下に導入する $\tilde{t}_n$ を用いる．このノイズモデルと拍節 HMM を組み合わせるため，ベルヌーイ分布に従う二値変数 $s_n\in\{\mathrm{S},\mathrm{N}\}$ を導入する（$P(s_n)=\alpha_{s_n}$; $\alpha_{\mathrm{S}}+\alpha_{\mathrm{N}}=1$）．$s_n=\mathrm{S}$ の場合，t_n は 4.1 節の拍節 HMM により生成される．$s_n=\mathrm{N}$ の場合，t_n は式 (4) により生成される．この確率過程は出力合流 HMM [18] により記述することができる．このモデルは状態変数は $z_n=(s_n,b_n,g_n,v_n,\tilde{t}_n)$ で表され，初期確率と遷移確率は以下で記述される（図 3）．

$$\begin{aligned}P(z_n|z_{n-1})=\ &\delta_{s_n\mathrm{N}}\,\alpha_{\mathrm{N}}\,\delta_{b_{n-1}b_n}\delta_{g_{n-1}g_n}\delta(v_n-v_{n-1})\delta(\tilde{t}_n-\tilde{t}_{n-1})\\ &+\delta_{s_n\mathrm{S}}\,\alpha_{\mathrm{S}}\,P(b_n,g_n|b_{n-1})P(v_n|v_{n-1})P(\tilde{t}_n|\tilde{t}_{n-1}),\end{aligned}\quad(5)$$

$$P(t_n|z_n)=\delta_{s_n\mathrm{S}}\,\delta(t_n-\tilde{t}_n)+\delta_{s_n\mathrm{N}}P_*(t_n|\tilde{t}_n).\quad(6)$$

ここで，δ は離散変数に対しては Kronecker のデルタ，連続変数に対しては Dirac のデルタ関数を表し，$P(\tilde{t}_n|\tilde{t}_{n-1})$ は式 (3) で与えられる．変数 $\tilde{t}_n$ はシグナルモデルの一つ前の発音時刻を記憶するために用いられる．$\alpha_{s_{n'}}=\mathrm{S}$ を満たす最大の $n'<n$ に対して，$\tilde{t}_n=t_{n'}$ となる．

音符トラックデータの音長とベロシティーの情報は，余分な音を認識する際に役立つと考えられる．これは，余分な音符のこれらの特徴量の分布は，楽譜由来の音符に対して一般的に平均と分散が小さいからである．この情報を用

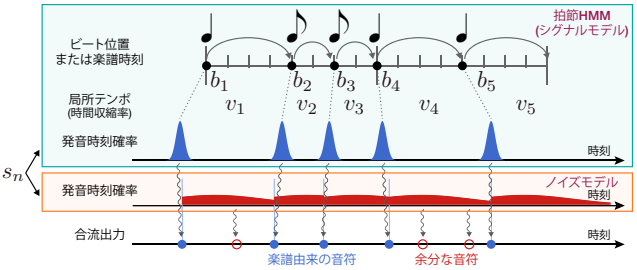

図 3　Noisy 拍節 HMM における発音時刻の生成過程

いるため，上記のモデルを各音符に対する特徴量 f_n の生成過程を含むものに拡張する．記法を簡単にするため，一般の特徴量に対して記号 f_n を用いるものとする．特徴量に対する分布は，変数 s_n に条件付いた以下の確率分布で定義する．

$$P(f_n = f) = \delta_{s_n\mathrm{S}} P(f|\mathrm{S}) + \delta_{s_n\mathrm{N}} P(f|\mathrm{N}). \qquad (7)$$

音長とベロシティーはどちらも正の数で定義されるため，ここでは $P(f|s) = \mathrm{IG}(f; a_s, b_s)$ と表されると仮定する（$\mathrm{IG}(x; a, b) = b^a x^{-a-1} e^{-b/x}/\Gamma(a)$ は形状パラメータ a とスケールパラメータ b を持つ逆ガンマ分布を表す）．特徴量の導入は，α_{s_n} に対する以下の変更として記述できる．

$$\alpha_{s_n} \to \alpha'_{s_n} = \alpha_{s_n} \prod_{f:\ \text{features}} P(f_n|s_n)^{w_f}. \qquad (8)$$

ここで，正規のモデルに対しては $w_f = 1$ である．導入する特徴量の数は恣意的であるため，w_f を変数として捉え，最尤法などにより最適化することが考えられる．本研究では，5 節に述べるように，w_f は採譜の誤り率によって最適化する．Noisy 拍節 HMM の推論アルゴリズムは，文献 [18] で開発されたテクニックを適用することにより導出できる．

5. 評価

5.1 評価尺度

3 節に述べた多重音検出部分の性能評価には，MIREX の評価でも用いられる，文献 [28] で導入されている発音ベースの音符トラッキングの評価尺度を用いる．この尺度では，音符が正解と同じ音高を持ち，発音時刻が正解の発音時刻に比べ ±50 ms 以内にあるとき正しく検出されたとみなす．この基準に基づき，適合率（precision）$\mathcal{P}_\mathrm{n}$，再現率（recall）$\mathcal{R}_\mathrm{n}$，F 値 $\mathcal{F}_\mathrm{n}$ を定義する．

正解楽譜との比較に基づき採譜結果の楽譜を評価する尺度は，リズム量子化の文脈で既に議論されている [18, 19]．発音楽譜時刻に関して採譜結果を修正するのに必要なスケールおよびシフト編集の最小数として定義されるリズム修正コスト（rhythm correction cost; RCC）[18] は，発音時刻誤り率として用いることができる．また，後続の発音楽譜時刻に対して相対的に定義される消音楽譜時刻の誤りをカウントして消音時刻誤り率が定義できる [19]．これらの尺度を余分な音符が含まれる場合にも拡張するために，まず推定楽譜と正解楽譜の間で音符マッチング [29] を行い，合致音符（音高誤りを含む）・余分な音符・不足音符を特定する．（同様のアイディアは文献 [21] でも議論されている．）

正解楽譜の音符数を N_GT，推定楽譜の音符数を N_est，音符誤りの数を N_p，余分な音符数を N_e，不足音符数を N_m，合致音符数を $N_\mathrm{match} = N_\mathrm{GT} - N_\mathrm{m} = N_\mathrm{est} - N_\mathrm{e}$ とする．これらを用いて，「音高誤り率」を $E_\mathrm{p} = N_\mathrm{p}/N_\mathrm{GT}$，「余分音符率」を $E_\mathrm{e} = N_\mathrm{e}/N_\mathrm{est}$，「不足音符率」を $E_\mathrm{m} = N_\mathrm{m}/N_\mathrm{GT}$，「発音時刻誤り率」を $E_\mathrm{on} = \mathrm{RCC}/N_\mathrm{match}$，「消音時刻誤り率」を $E_\mathrm{off} = N_\mathrm{o.e.}/N_\mathrm{match}$ で定義する．ここで RCC の計算は文献 [18] に説明されており，$N_\mathrm{o.e.}$ は文献 [19] と同様に最接の発音楽譜時刻で正規化した消音楽譜時刻が誤っている音符の数である．これらの 5 つの尺度の平均として「総合誤り率」E_all を定義する．

5.2 実験設定

3 節の音響モデルの学習には，MAPS データベース [22] の単一音の録音データから抽出されたスペクトルテンプレートの辞書を用いる．この辞書には，評価データとして用いる ‘ENSTDkCl’ モデル以外の 8 種類のピアノに対するものが含まれている．音高の範囲としてはピアノの音域（A0 から C8）を用いる．4 節の記号処理のモデルのパラメータの内，$P(b_1, g_1)$, $P(b_n, g_n|b_{n-1})$, v_ini, $\sigma_{\mathrm{ini}\,v}$, σ_v は従来研究 [18] と同じものを用い，α_s, a_s, b_s は多重音検出手法の出力から学習した結果を用いる．この他のパラメータ σ_*, σ_t, λ_t, w_f は E_all の最大化によりテストデータを用いて最適化する．

採譜システムの評価には，MAPS データベース [22] の ‘ENSTDkCl’ セットに含まれる 30 のピアノ録音とそれらに対応する正解音符トラックデータと MusicXML 形式の楽譜データを用いる．多重音検出の既存研究と同様に，それぞれの録音の最初の 30 s を評価の対象とする．比較のため，多重音検出手法である調和 NMF（harmonic NMF; HNMF）も評価する．これは，ナローバンドスペクトルの重み付き和として音高ごとのスペクトルをモデルかする適応型 NMF に基づく手法である．リズム量子化部分に関しては，本研究の手法を組み合わせる．

5.3 結果

表 1 に多重音検出手法の精度を示す．文献 [7] の PLCA 手法を PLCA-4D，3.2 節での音符トラッキングの改良を行ったものを PLCA-4D-NT を表す．PLCA-4D-NT は，PLCA-4D に対して音符ベース F 値で約 1%上回っており．より低い適合率とより高い再現率を示した．PLCA-4D-NT での高い再現率は，余分な音符は除去できるが不足音符を

Method	$\mathcal{P}_n$	$\mathcal{R}_n$	$\mathcal{F}_n$	p-value
HNMF [5]	62.3	76.9	67.9	0.0034
PLCA-4D [7]	79.4	66.0	71.7	0.080
PLCA-4D-NT	77.9	68.9	72.8	—

表 1 MAPS-ENSTDkCl データセットでの多重音検出結果の平均精度 (%) に関する音響モデルの比較．最後の列は PLCA-4D-NT に対する $\mathcal{F}_n$ の p 値を示す．

Method	E_p	E_m	E_e	E_{on}	E_{off}	E_{all}	p-value
Finale 2014	5.6	24.2	18.3	53.3	54.0	31.1	$< 10^{-5}$
MuseScore 2	6.1	26.1	16.9	39.7	56.3	29.0	$< 10^{-5}$
拍節 HMM-def	4.8	25.2	15.7	29.6	41.9	23.5	0.023
拍節 HMM	4.7	25.4	16.3	23.6	40.9	22.2	0.18
Noisy 拍節 HMM	4.4	28.6	13.3	21.6	39.3	21.4	—

表 2 MAPS-ENSTDkCl データセットでの採譜結果の平均誤り率 (%) に関するリズム量子化手法の比較．PLCA-4D-NT の出力音符トラックデータに対する結果を示す．最後の列は Noisy 拍節 HMM に対する E_{all} の p 値を示す．

Method	E_p	E_m	E_e	E_{on}	E_{off}	E_{all}	p-value
Finale 2014	10.7	18.3	39.3	57.2	57.4	36.6	$< 10^{-5}$
MuseScore 2	12.3	19.9	34.4	49.7	62.6	35.8	$< 10^{-5}$
拍節 HMM-def	10.5	18.6	33.2	36.5	44.1	28.6	$< 10^{-5}$
拍節 HMM	9.6	17.5	33.0	25.5	42.1	25.5	0.00048
Noisy 拍節 HMM	7.2	20.8	19.8	24.1	41.2	22.6	—

表 3 HNMF [5] の出力音符トラックデータに対する表 2 と同様の結果．

補うことはできない Noisy 拍節 HMM での入力としてより適していると考えられる．HMNF [5] は，最も高い再現率を示したが，F 値は最も低かった．

表 2 と 3 に採譜システム全体の評価結果を示す．比較のため，演奏 MIDI データのリズム量子化を扱った先行研究 [18] のパラメータを用いた拍節 HMM（拍節 HMM-def と記す）の結果およびパラメータ最適化をした拍節 HMM と Noisy 拍節 HMM の結果を記す．また音符トラックデータを 2 種類の楽譜ソフト（MuseScore 2 [23] と Finale 2014 [24]）により MusicXML 形式に変換した結果も比較する．PLCA-4D-NT と HNMF の両方の出力に対して，Noisy 拍節 HMM が総合誤り率の平均で他を上回り，楽譜ソフトの結果に比べ優位に低い値を示したことが確認できる．拍節 HMM のパラメータの最適化が誤り率の低下に有効であることも確認できる．拍節 HMM に比べ，Noisy 拍節 HMM は E_m 以外の全ての誤り率を低下することができており，その効果は適合率が高い HNMF の出力結果に対してより顕著である．図 4 に示す例では，Noisy 拍節 HMM が余分な音符（10.23 s にある G4）を正しく削除し，拍節 HMM-def でおこる和音の認識誤り (第 4 小節の E♭4 と G4) を修正できていることが確認できる．

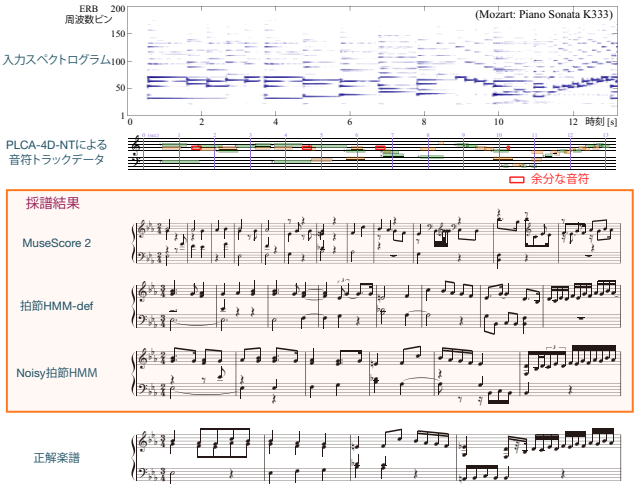

図 4 採譜結果の例（MAPS-ENSTDkCl データセット，Mozart: Piano Sonata K333）．

6. 結論

本稿では，多声音楽採譜のための多重音検出とリズム量子化手法の統合について述べた．PLCA を用いた多重音検出手法を音符トラッキングに関して改良した手法と Noisy 拍節 HMM に基づく音符トラックデータの余分な音符を低減できるリズム量子化手法を提案し，それぞれが採譜精度の向上に有効であることを確認した．また，演奏の時間揺らぎを表す拍節 HMM のパラメータの最適化が採譜の誤りを減らす上で有効であることも確認した．

現時点では，音楽的・音響的に簡単な場合を除いて，提案システムによる採譜結果は音楽的に不適切な音高配置や演奏不可能な音符などが含まれており，改善の余地が大きい．現在の Noisy 拍節 HMM は音高情報を記述していないが，音高のモデルを取り込むことで，不自然な音高を持つ音符を低減できるようになると考えられる．音符トラックデータの誤り音符の内，音高誤りや不足音符など余分な音符以外のものを修正することは現状ではできていない．このためには記号的音楽言語モデルと音響モデルの統合が必要だと考えられる．主観評価を含めた，より徹底した評価は今後の課題である．また評価尺度の計算における，楽譜同士のマッチングの誤りの影響も調べることも課題である．

謝辞 本研究は，科研費 24220006, 26280089, 26700020, 15K16054, 16H01744, 16H02917, 16K00501, 16J05486, JST ACCEL No. JPMJAC1602 からの支援を受けた．中村は，日本学術振興会の特別研究員制度および電子通信普及財団の長期海外研究滞在支援制度からの支援を受けた．

参考文献

[1] A. Klapuri and M. Davy (eds.), *Signal Processing Methods for Music Transcription*, Springer, 2006.
[2] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, “Automatic music transcription: Challenges

and future directions," *J. Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

[3] S. Levinson, L. Rabiner, and M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *The Bell Sys. Tech. J.*, vol. 62, no. 4, pp. 1035–1074, 1983.

[4] C. Raphael, "A graphical model for recognizing sung melodies," in *Proc. ISMIR*, 2005, pp. 658–663.

[5] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE TASLP*, vol. 18, no. 3, pp. 528–537, 2010.

[6] K. O'Hanlon and M. D. Plumbley, "Polyphonic piano transcription using non-negative matrix factorisation with group sparsity," in *Proc. ICASSP*, 2014, pp. 3112–3116.

[7] E. Benetos and T. Weyde, "An efficient temporally-constrained probabilistic model for multiple-instrument music transcription," in *Proc. ISMIR*, 2015, pp. 701–707.

[8] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM TASLP*, vol. 24, no. 5, pp. 927–939, 2016.

[9] R. Kelz, M. Dorfer, F. Korzeniowski, S. Böck, A. Arzt, and G. Widmer, "On the potential of simple framewise approaches to piano transcription," in *Proc. ISMIR*, 2016, pp. 475–481.

[10] H. Longuet-Higgins, *Mental Processes: Studies in Cognitive Science*, MIT Press, 1987.

[11] D. Temperley and D. Sleator, "Modeling meter and harmony: A preference-rule approach," *Comp. Mus. J.*, vol. 23, no. 1, pp. 10–27, 1999.

[12] A. T. Cemgil, P. Desain, and B. Kappen, "Rhythm quantization for transcription," *Comp. Mus. J.*, vol. 24, no. 2, pp. 60–76, 2000.

[13] C. Raphael, "A hybrid graphical model for rhythmic parsing," *Artificial Intelligence*, vol. 137, pp. 217–238, 2002.

[14] M. Hamanaka, M. Goto, H. Asoh, and N. Otsu, "A learning-based quantization: Unsupervised estimation of the model parameters," in *Proc. ICMC*, 2003, pp. 369–372.

[15] H. Takeda, T. Otsuki, N. Saito, M. Nakai, H. Shimodaira, and S. Sagayama, "Hidden Markov model for automatic transcription of MIDI signals," in *Proc. MMSP*, 2002, pp. 428–431.

[16] D. Temperley, "A unified probabilistic model for polyphonic music analysis," *J. New Music Res.*, vol. 38, no. 1, pp. 3–18, 2009.

[17] A. Cogliati, D. Temperley, and Z. Duan, "Transcribing human piano performances into music notation," in *Proc. ISMIR*, 2016, pp. 758–764.

[18] E. Nakamura, K. Yoshii, and S. Sagayama, "Rhythm transcription of polyphonic piano music based on merged-output HMM for multiple voices," *IEEE/ACM TASLP*, vol. 25, no. 4, pp. 794–806, 2017.

[19] E. Nakamura, K. Yoshii, and S. Dixon, "Note value recognition for piano transcription using Markov random fields," *IEEE/ACM TASLP*, vol. 25, no. 9, pp. 1542–1554, 2017.

[20] E. Kapanci and A. Pfeffer, "Signal-to-score music transcription using graphical models," in *Proc. IJCAI*, 2005, pp. 758–765.

[21] A. Cogliati and Z. Duan, "A metric for music notation transcription accuracy," in *Proc. ISMIR*, 2017, pp. 407–413.

[22] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE TASLP*, vol. 18, no. 6, pp. 1643–1654, 2010.

[23] MuseScore, "MuseScore 2," https://musescore.org/en [online], accessed on: Oct. 11, 2017.

[24] MakeMusic, "Finale 2014," https://www.finalemusic.com [online], accessed on: Oct. 11, 2017.

[25] E. Nakamura, N. Ono, and S. Sagayama, "Merged-output HMM for piano fingering of both hands," in *Proc. ISMIR*, 2014, pp. 531–536.

[26] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizations," *Computational Intelligence and Neuroscience*, 2008, Article ID 947438.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[28] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of multiple-F0 estimation and tracking systems," in *Proc. ISMIR*, 2009, pp. 315–320.

[29] E. Nakamura, K. Yoshii, and H. Katayose, "Performance error detection and post-processing for fast and accurate symbolic music alignment," in *Proc. ISMIR*, 2017, pp. 347–353.