統計的音楽理論と演奏モデルとベイズ文法推論の融合に基づく自動編曲手法の確立

代表研究者　　中村　栄太　　京都大学　大学院情報学研究科　特別研究員(PD)

1 はじめに

近年、音楽コンテンツはインターネットをはじめとする電気通信における一大コンテンツとなっており、様々な楽曲がインターネットを通じて配信され、ユーザが自ら作曲、編曲、演奏した多くの動画・音声が共有されている。一方で、計算機による音楽の情報処理技術の研究が盛んになってきており、音楽演奏信号からユーザにとって演奏可能な楽譜情報に変換する採譜や既存の音楽のスタイルや楽器編成を変更して、より多様な音楽の楽しみ方を可能とする編曲など、従来は音楽専門家の手でのみ可能であった技能を自動化した技術の研究開発が広く行われている。

本研究は、統計機械学習手法に基づいた自動音楽編曲（ある音楽を別の音楽スタイルや楽器編成へと転換）の基礎技術の開発およびその際に必要となる自動採譜技術（音楽音響信号を楽譜へと転換）の開発を目標としている。以下の３つの主要なテーマに取り組んだ。

・ピアノ採譜における音価認識
・音響モデルと記号モデルの統合によるピアノ採譜
・確率的音型モデルの構築

採譜や編曲の定式化では、音楽に内在する「文法構造」や音楽演奏に含まれる時間のゆらぎや演奏誤りなどを如何にモデルで捉えるかが問題となる。こうした、楽譜モデルおよび演奏モデルの構築および学習・推論手法の確立が、技術的に共通した課題となっている。各テーマについては次節以降に詳しく記す。

音楽データは、音響データと楽譜を表す記号データに大きく分けられる。例えば、自動採譜は音楽音響データを記号データへと変換する課題である。よって、自動採譜や自動編曲技術の確立には、音楽言語モデルと音響モデルの統合は最重要課題の一つである。研究代表者はこれまで、主に記号データの情報処理のための統計的手法の研究に取り組んできた。一方で、滞在先のクイーン・メアリー・ロンドン大学の Simon Dixon 教授および Emmanouil Benetos 助教らは音楽音響処理の専門家であり、本共同研究により両者の専門知識を融合したテーマに取り組むことができた。本研究の成果を拡張して、今後さらに音楽情報処理技術の進展が期待できるが、これに関しては最後の節で議論する。

2 ピアノ採譜における音価認識

2-1 背景

従来から自動採譜に関する研究は多いが、ピアノや合奏など同時に複数の音が存在する音楽（**多声音楽**）の採譜は依然として未解決問題であり、多くの取り組みがなされている[1]。多声音楽では、各時刻で音響的変動が複雑に重ね合わさるため、信号処理のみによる解決は難しいと考えられており、楽譜の事前知識を記述した**楽譜モデル**を用いるアプローチが有効と考えらえる。また、多声音楽の採譜を単純化した問題として、MIDI 情報（各演奏音符の音高と発音時刻、消音時刻、音量のリスト）で表されるピアノ演奏の採譜（**リズム量子化**とも呼ばれる）を考え、楽譜モデルと演奏モデルの統合による採譜手法が研究されてきたが、音高と発音時刻のみ扱う研究がほとんどで、楽譜表示で重要である消音時刻あるいは音価（楽譜上での音長）を扱う研究は少ない。

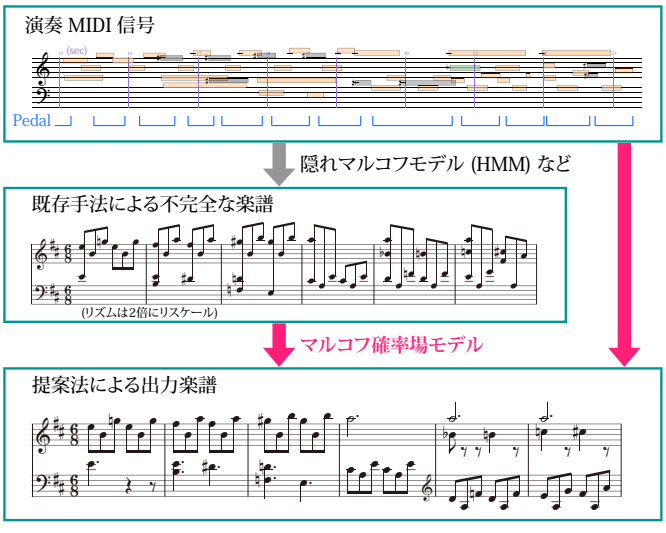

図1 マルコフ確率場モデルを用いた音価認識の概略

発音時刻に関するリズム量子化については既に高精度の手法があることを踏まえ[2]、本研究では、音高と発音時刻、そして元の MIDI 演奏のタイミング情報が得られる状況で、音価を推定する問題を考え、統計モデルに基づく手法を提案する（図 1）。まず、実際の楽譜の音価に関する統計解析を行い、発音楽譜時刻と音高の文脈、そして音価同士の相互依存性が音価の予測モデル構築の鍵となることを示す。次に、音価の事前分布を与える楽譜モデルと演奏音符の音長の依存関係を記述する演奏モデルの統合したマルコフ確率場を構成する。楽譜モデルで用いる文脈あるいは特徴量として最適なものを決定するため、文脈木クラスタリング法に基づく手法を開発する。最後に、実際のデータで提案法を用いた音価認識を行い、系統的な評価をする。

2-2 演奏における音長および楽譜における音価の統計解析

まず、演奏における音長と対応する楽譜の音価の関係を調べる。ピアノ演奏ではペダルの有無により音響的な音長が変化するため、ペダルには依存しない鍵盤の離鍵のタイミングによって決まる**鍵保持時間**とペダルの有無まで含めた**ダンパー上昇時間**の 2 種類を考える。これらの音長と演奏の局所テンポと音価によって期待される理想的な音長の比の分布を図 2 に示す。どちらの分布も大きな標準偏差を持っていることから、単純に演奏音符の音長から楽譜の音価を推定することは難しいことが確かめられる。これにより、高精度な音価の推定には、楽譜内の音価の配置の事前分布を記述する楽譜モデルが必要であることがわかる。

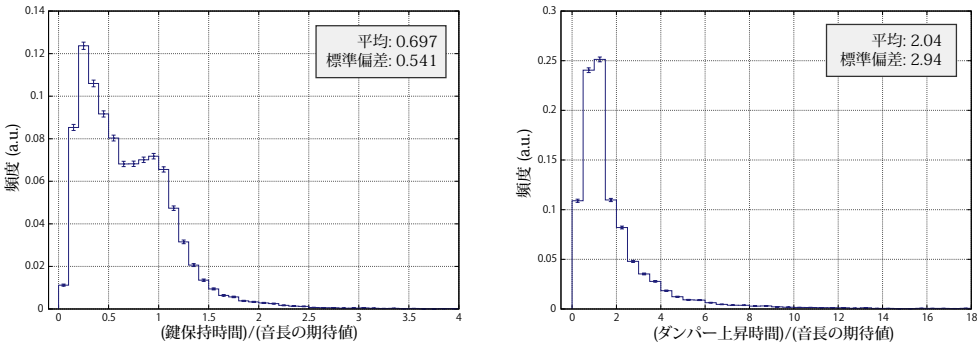

図2 実ピアノ演奏データにおける鍵保持時間とダンパー上昇時間の分布

通常の楽譜では、ある音符の消音楽譜時刻は、別の音符の発音楽譜時刻に多くの場合、一致することが経験的に知られている。実際、図 3 に示すように、ある音符に対して、その後続の音符の発音時刻との差により定義される音価を IONV（発音間音価）と呼ぶとき、何番目の IONV と音価が一致するか（あるいは一致しないか）に関する分布は図 4(a)のようになる。さらに、この IONV 空間上での実際の音価の分布と対応する音符の音高の文脈の依存性を調べるため、特定の音高の文脈での分布を図 4(b)(c)に示す。図 4(b)は一つ後ろの和音に 5 半音以内の近さの音高が存在する音符のみに対する分布であり、図 4(c)は一つ後ろの和音には 14 半音以内の近さの音高はなく、二つ後ろの和音には 5 半音以内の近さの音高が存在する音符に対する分布である。この例が示すように、各音符に対してその音高の文脈に依存した音価の分布を考えることにより、予測性能の高い楽譜モデルを構成できると考えられる。

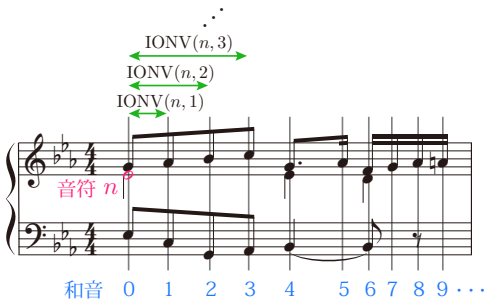

図3 IONV(発音間音価)の例

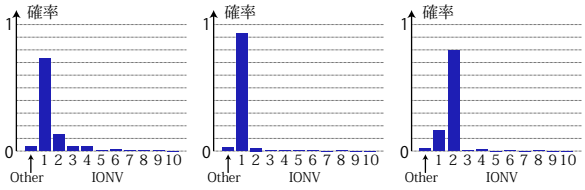

図4 IONVと比較した音価の分布

2-3 提案モデル

音価の事前分布である楽譜モデルと演奏モデルを統合するため、以下のマルコフ確率場を考える。

$$P(\boldsymbol{r}|\boldsymbol{p},\boldsymbol{d},\bar{\boldsymbol{d}},\boldsymbol{\tau},\boldsymbol{v})$$

$$\propto \exp\Bigg[-\sum_{n=1}^{N} H_1(r_n;\boldsymbol{\tau},\boldsymbol{p}) - \sum_{(n,m)\in\mathcal{N}} H_2(r_n,r_m) - \sum_{n=1}^{N} H_3(r_n;d_n,\bar{d}_n,v_n)\Bigg].$$

$$H_1 = -\beta_1 \ln P(r_n;\boldsymbol{\tau},\boldsymbol{p}), \quad H_2 = -\beta_2 \ln P(r_n,r_m),$$

$$H_3 = -\beta_{31}\ln P(d_n;r_n,v_n) - \beta_{32}\ln P(\bar{d}_n;r_n,v_n).$$

ここで、r_n, p_n, d_n, d_n, τ_n, v_n は n 番目の音符の音価、音高、鍵保持時間、ダンパー上昇時間、発音楽譜時刻、局所テンポを表し、β_1, β_2, β_{31}, β_{32} は正の定数である。3 つの項 H_1, H_2, H_3 はそれぞれ音高の文脈に依存する**文脈モデル**、近接する音価どうしの**相互依存モデル**、そして**演奏モデル**を表す（図 5）。この内、H_2 と H_3 は、直接データを用いた学習により構成できるが、H_1 は音高の文脈が組み合わせ爆発を起こすため単純には計算可能なモデルとはならない。本研究では簡単のため、後続する 10 個の和音のそれぞれについて半音単位で最接の音高のリストを音高の文脈として用いる。この文脈に対して文脈木クラスタリング[3]を適用することにより、計算可能な文脈モデルを構成する。

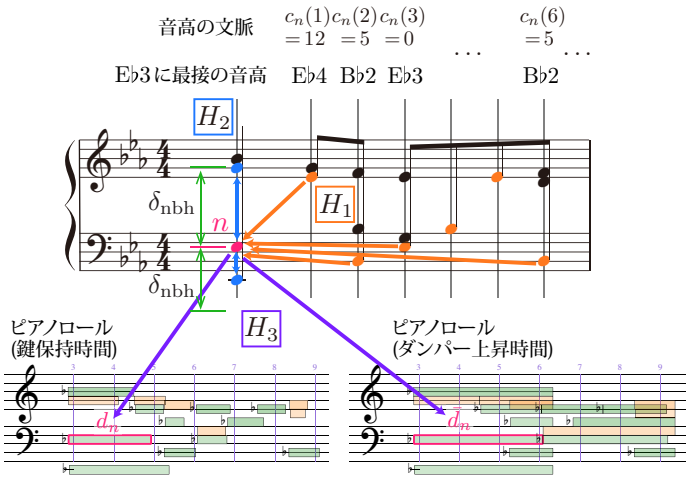

図5 音価認識のためのマルコフ確率場モデル

図 6 に文脈木クラスタリング学習の結果の一部を示す。図中の各分布において、1 から 10 は 1 番目から 10 番目の IONV に対する音価の確率、0 はその他の確率を示しており、各ノードの下には対応する文脈の条件が記されている（$c(n)$は後続の n 番目の和音における最接の音高との半音単位での距離を示す）。また、各ノードの上にはノードの ID と対応するデータサンプル数とその割合が、青い枠の内側には分布の最大確率値が記されている。例えばノード[6]や[9]を見ると、2 番目や 3 番目の IONV の確率が高くなっているが、これらのノードは 2 番目や 3 番目の後続和音に音高が近い音符が存在する条件で定義されている。文脈木クラスタリングの学習結果は、楽譜における声部構造を反映したものになっていることが分かる。

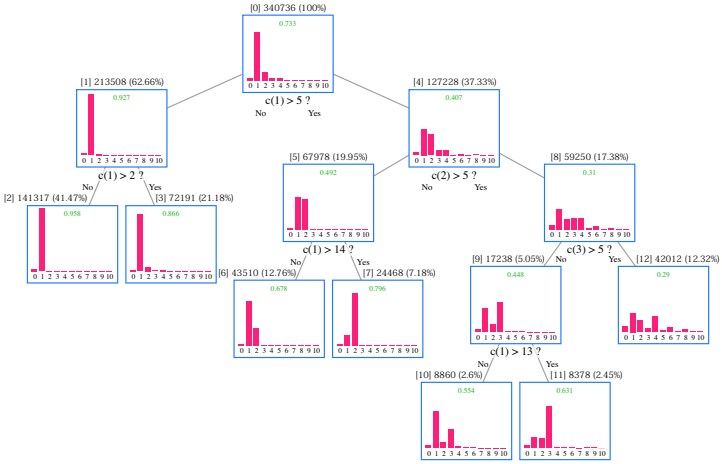

図6 文脈クラスタリングの学習結果

2-4 評価結果

前節に述べた提案モデルと他の手法の実データを用いた評価について説明する。まず提案モデルについては、拍節 HMM（隠れマルコフモデル）を用いて発音時刻のリズム量子化を行った結果に対してフルモデルを適用した場合と、相互依存モデルおよび演奏モデルを取り除いた場合のモデルを適用した場合を評価した。さらに、フルモデルで楽譜モデルに 1 リーフの文脈木を用いた場合と演奏モデルのみと音価のユニグラム分布を用いた場合も比較評価した。また、リズム量子化の先行研究として知られる Melisma Analyzer [4]を用いて発音時刻と音価の両方を推定した結果、そして発音時刻のみ Melisma Analyzer を用いて推定して、その後提案モデルを用いて音価を推定した結果も比較評価した。評価尺度には、推定された音価を各音符の 1 番目の IONV で規格化した相対音価を正解データの対応する相対音価と比較した際の誤り率を用いる。評価データには独自に収集したクラシックのピアノ演奏データ（60 フレーズ×3 演奏者）を用いた。

結果を図 7 に示す。まず従来手法の Melisma Analyzer では 70%近くある誤り率が提案法によって約 26%にまで大幅に低減されたことが分かる。また Melisma Analyzer の発音時刻の推定結果に提案法を適用した場

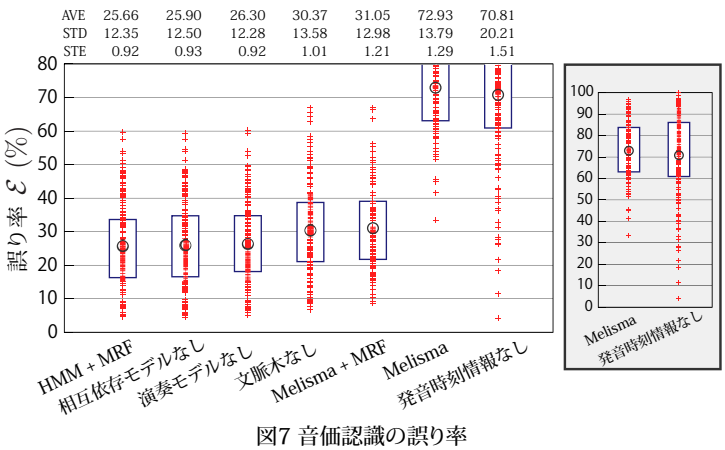

図7 音価認識の誤り率

合でも、誤り率が 31%程度であることにより、提案法による誤り低減は、特定の発音時刻の推定手法に寄らないことも確認された。また提案法における異なる条件下での結果を比べると、演奏モデルを用いない場合でも、誤り率は大きく変化しないことが分かる。またそれに比べて、楽譜モデルにおける文脈木クラスタリングの重要性が高いことも確認出来る。以上より、音価推定においては、演奏音符の音長や演奏モデルの影響は少なく、楽譜モデルが高精度な推定に欠かせないこと、そして提案モデルにより現時点では最高精度の音価推定が行えることが確認できた。

また提案法を用いたピアノ MIDI 演奏の採譜結果の例を図 8 に示す。若干の推定ミス（赤色で表示）はあるものの、右手のメロディーと左手の和音伴奏の構造を捉えた採譜結果になっていることが確認できる。

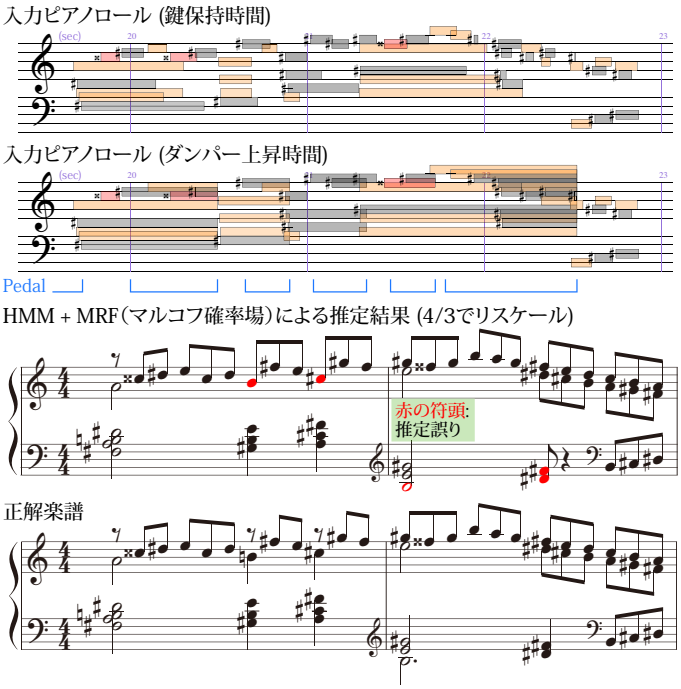

図8 提案法による採譜結果の例

2-5 まとめ

自動採譜において、最終的に楽譜を得る上で欠かせない音価の推定手法を提案した。提案手法の誤り率は約 26%であり、今後改善の余地は大きいが、現時点では最高精度を達成したと言える。また統計解析および提案モデルを用いた評価の結果から、音価推定では他の音符の音高と発音時刻の情報を基にした楽譜モデルを用いることが有効であり、演奏音符の音長や演奏モデルの影響は小さいことが分かった。これは今後、モデルや手法を改良する上で重要な指標の一つになると考えられる。

以上の内容および研究内容の詳細は以下の発表論文に記載されている。

・E. Nakamura, K. Yoshii, and S. Dixon, “Note Value Recognition for Piano Transcription Using Markov Random Fields,” IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 25, no. 7, pp. 1542-1554, 2017.

3 音響モデルと記号モデルの統合によるピアノ採譜

3-1 背景

自動採譜の最終目標の一つは、音楽音響信号を人間が読める楽譜データに変換することである。このためには楽譜情報である音高およびリズムの量子化が必要である。音声認識と同様に、音楽楽譜モデルと音響モデルの統合によるアプローチが考えられるが、ピアノや合奏曲などの多声音楽の場合は、和音の音高配置の可能性が膨大な数になるなど、計算量的に現実的な手法は現在のところ実現できていない。そこで、これまで多声音楽の自動採譜問題は、音響信号から量子化された音高を持つ音符列へと変換する**多重音検出**とそうした音符列に対して発音および消音楽譜時刻の認識を行う**リズム量子化**手法に分けて研究されてきた。これらの問題に関する研究は近年大きく進展している一方で、別々に研究されており、これまで両者の統合

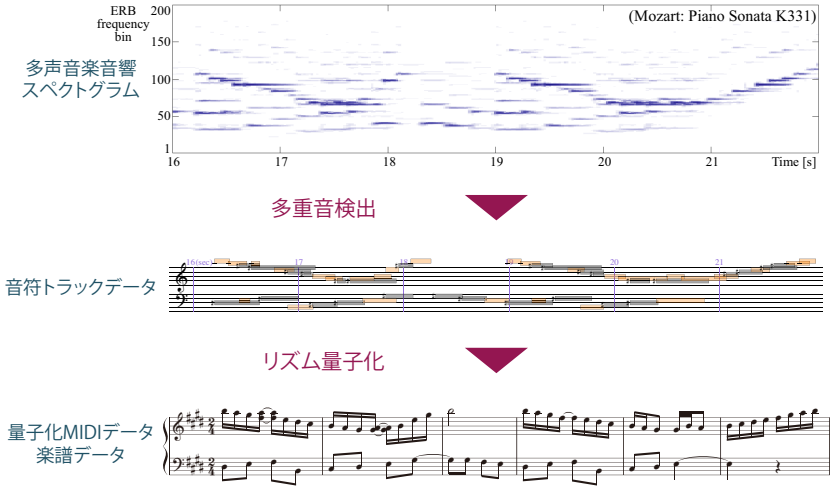

図9 多重音検出とリズム量子化の統合による多声音楽の自動採譜

による完全な自動採譜システムの構築は実現されていない。

本研究では、最新の多重音検出手法とリズム量子化手法の統合による完全な自動採譜システムの構築を行い、系統的な評価を行う（図 9）。多重音検出手法では、後続のリズム量子化に適した出力が得られるように音符トラッキング部の改良を行う。またリズム量子化手法では、前段の多重音検出手法で生じる誤り音符を削減できる機構として、新たに**ノイズ拍節 HMM** に基づく手法を提案する。また評価のための尺度も構成し、手法の改良の有効性を検証する。

3-2 システム構成

まずシステム全体の構成について説明する（図 10）。前述の通り、まず音楽音響信号に対して多重音検出を適用して、音高（半音単位）と発音・消音時刻（秒単位）および音量の情報を持つ音符列データを得る（これを**音符トラックデータ**と呼ぶ）。この過程では、まず時間フレームごとに音高の有無の情報を推定する多重音解析を適用し、その情報を元に音符の検出を行う音符トラッキングを適用する。

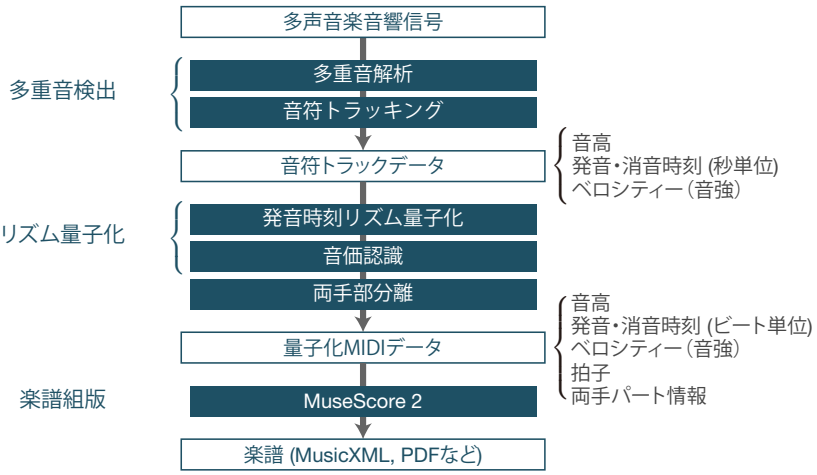

図10 システム構成

次に音符トラックデータに対して、リズム量子化手法を適用して、各音符に対してビート単位での発音・消音楽譜時刻を推定し、さらに拍子の推定を行う。この過程では、まず音符のオンセットのみを用いた発音時刻の量子化およびテンポトラッキングを行い、次にその情報を元に消音楽譜時刻の推定を行う。なお消音楽譜時刻の推定には 2 節で説明したマルコフ確率場に基づく音価認識手法を適用する。得られた音符列データに対して、出力合流 HMM に基づく右手左手パートの分離[5]を行い、量子化 MIDI データを得る。

最終的に人間が読める楽譜データ（PDF 形式や MusicXML 形式）を得るために、量子化 MIDI データに対して楽譜編集ソフトである MuseScore 2 の MIDI 入力機能を用いる。MuseScore 2 では右手左手パートそれぞれの中の声部分離を自動で行える。

3-3 多重音検出

まず時間フレームごとに音高の有無の情報を推定する**多重音解析**について説明する。本研究では PLCA（probabilistic latent component analysis）に基づく手法を用いる[6]。まず入力音響信号に対して、ERB（equivalent rectangular bandwidth）法によりパワースペクトログラムを得て、これを規格化したものを同時確率分布 $P(f,t)$と見なす（f は周波数ビン、t は時間フレームを表す）。PLCA ではこれをベイズの公式を用いて以下のように分解することを考える。

$$P(f,t) = P(t) \sum_{q,p,i} P(f|q,p,i) P_t(i|p) P_t(p) P_t(q|p)$$

右辺において、p は音高、q は音状態（ピアノではアタック・サステイン・リリースに対応）、i はピアノの種類に対応する添字であり、$P(f|q,p,i)$は周波数スペクトルのテンプレートを表す。右辺への各要素は EM アルゴリズムを用いて推定可能である。これにより得られた確率 $P(t,p)=P(t)P_t(p)$は各時間フレームでの音高の分布を表しており、これが多重音解析の出力となる。

次に音符トラッキングにおいては、まず多重音解析の結果に閾値処理をして、30 ms 未満の音長の音符を取り除いたものを第一の音符トラックデータとする。次に、この結果に対して元のパワースペクトルにピーク検出を適用した情報を用いて、繰り返し音の検出を行う。さらに各オンセット時刻の精密な値を求めるため、前後 50 ms の領域においてパワースペクトルの変化によりマッチングを行う。これにより最終的な音符トラックデータを得る。

3-4 リズム量子化

ここでは、音符トラックデータを元に発音時間の量子化とテンポトラッキングを行うオンセットリズム量子化の手法について説明する。従来法として拍節 HMM を用いた方法が知られている[7]。これは、発音楽譜

時刻を小節内のビート位置として表し、この空間上でのマルコフ過程により記述される楽譜モデルと、局所的なテンポを潜在状態空間として持つ、ガウシアン・マルコフ過程として記述される演奏モデルを統合したモデルである。この方法では、全ての演奏音符を楽譜上の音符にマップすることを考えているが、多重音検出により得られる音符トラックデータには、誤検出による楽譜上の音符以外の誤り音符が含まれている。この誤り音符をリズム量子化と同時に検出できる方法として、拍節 HMM の拡張モデルについて以下記す。

誤り音符の検出をするため、演奏音符に加えて誤り音符が生成され合流する過程を記述するモデルを構成する（図 11）。演奏音符は拍節 HMM により生成されるため、楽譜に内在する周期性を反映したタイミングで演奏音符が生成される。これに対して、誤り音符は時間的に一様に近い分布で演奏音符が生成されるとする。これらの生成過程は出力合流 HMM として混合することができる。

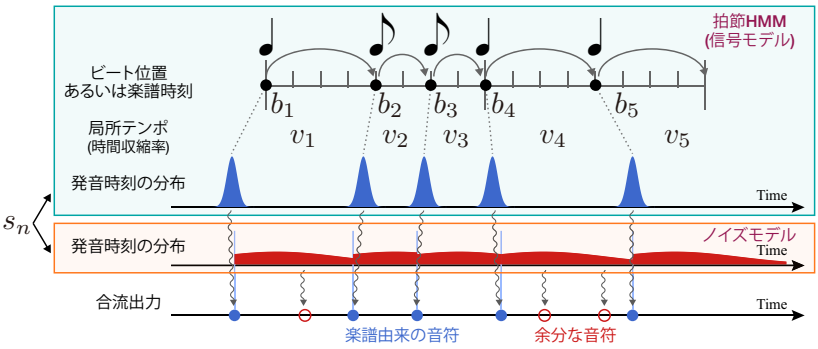

図11 ノイズ拍節HMMによる音符生成

実際の誤り音符を解析すると、その音長と音量は演奏音符に比べて小さい傾向があることが分かった。これによりこれらの特徴量を用いることで、より正確に誤り音符の認識が行えると考えられる。そこで上のモデルを拡張し、音長と音量に関する出力分布もモデルに取り込むことにする。以上のモデルは出力合流 HMM の推論手法で Viterbi 推定を行うことができる。

3-5 評価

まず改良した多重音検出手法の有効性を調べる。以下では、テストデータとして MAPS データベース内の「ENSTDkCl」ラベルのデータであるクラシックピアノ曲 30 曲を用いる。先行研究の結果との比較を可能とするため、各演奏の最初の 30 秒に関しての評価をする。音符トラックデータの評価に関しては、従来から広く用いられるオンセット基準の音符単位の F measure を用いる。この評価基準の計算では、正解データと比べ、推定の音符トラックデータに正解の音高と一致して、発音時刻のずれが 50 ms 以内である音符があるかどうかの条件で、Precision P_{n}、Recall R_{n}、F measure F_{n} を定義する。

結果を右表に示す。比較手法として、NMF（non-negative matrix factorization）ベースの手法として良く知られる HNMF（harmonic NMF）[8]と改良をする前の PLCA [6]の評価結果も示している。この結果により、提案手法の PLCA-4D-NT は HNMF および PLCA の従来法より高い F measure を達成していることが確認できる。また特に HNMF に比べ提案法では、Precision が高く、Recall が低い傾向が見られる。

Method	$\mathcal{P}_{\mathrm{n}}$	$\mathcal{R}_{\mathrm{n}}$	$\mathcal{F}_{\mathrm{n}}$	p-val.
HNMF	62.3	76.9	67.9	0.0034
PLCA-4D	79.4	66.0	71.7	0.080
PLCA-4D-NT	77.9	68.9	72.8	—

次に PLCA-4D-NT および HNMF から出力される音符トラックデータに対して、リズム量子化手法を適用して得られる採譜結果の評価について述べる。最終的な採譜結果の評価は、推定楽譜と正解楽譜を比較することで行う。この際、まず 2 つの楽譜同士でアラインメントを行うことにより、音符単位での不一致（脱落・挿入・置換誤り）として**不足音符率** E_{m}、**余分な音符率** E_{e}、**音高誤り率** E_{p} を定義する。次にアラインメントで得られたマッチした音符列（音高誤りを含む）を比べ、リズムの誤り率を求める。発音楽譜時刻の誤りに関しては、先行研究で提案されている音符単位のリズム変更と一定区間のスケール変換によるリズム変更の組み合わせにより定義される**リズム修正率** E_{on} を用いる[2]。消音楽譜時刻あるいは音価に関しては、後続の和音との IONV で規格化した相対音価を比較して、**音価誤り率** E_{off} を定義する。以上の 5 つの誤り率を推定楽譜の評価基準とする。またそれらの算術平均を**平均誤り率** E_{all} として、パラメータの最適化の際などに用いる。

採譜の評価結果を表に示す。表中で、MetHMM-def は従来の拍節 HMM をそのまま用いた結果、MetHMM は、拍節 HMM のパラメータを最適化したも

Method	E_{p}	E_{m}	E_{e}	E_{on}	E_{off}	E_{all}	p-val.
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	$< 10^{-5}$
MetHMM-def	4.8	25.2	15.7	29.6	41.9	23.5	0.023
MetHMM	4.7	25.4	16.3	23.6	40.9	22.2	0.18
NMetHMM	4.4	28.6	13.3	21.6	39.3	21.4	—

の、NMetHMM は提案法であるノイズ拍節 HMM を用いた結果を表す。また Finale 2014 および MuseScore 2 は既存の公開ソフトを音符トラックデータに直接適用した結果を表している。この結果より、本研究で構成した採譜システムの精度は公開ソフトの精度を大きく上回っていることが分かる。また平均誤り率は、NMetHMM が最も低く、次に MetHMM が低く、MetHMM-def が続いていることが分かる。これより、誤り音符の除去ができるノイズ拍節 HMM が採譜には有効であることが確かめられた。また拍節 HMM のパラメータの最適化の重要性も確かめられた。これは演奏を電子的に録音した MIDI 信号と多重音検出の出力結果である音符トラックデータではタイミングの逸脱の大きさが異なるためだと考えられる。

Method	E_{p}	E_{m}	E_{e}	E_{on}	E_{off}	E_{all}	p-val.
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	$< 10^{-5}$
MetHMM-def	10.5	18.6	33.2	36.5	44.1	28.6	$< 10^{-5}$
MetHMM	9.6	17.5	33.0	25.5	42.1	25.5	0.00048
NMetHMM	7.2	20.8	19.8	24.1	41.2	22.6	—

実際の採譜結果の一例を図 12 に示す。この例では、NMetHMM は誤り音符を一つ正しく認識しており、また MetHMM-def の結果で第 3 小節に見られる和音のミスアラインメントを修正することができた。この様な比較的簡単な楽曲の場合は、誤りはあるものの概ね再現性が高くかつ演奏可能な楽譜が採譜できていることが確認できる。

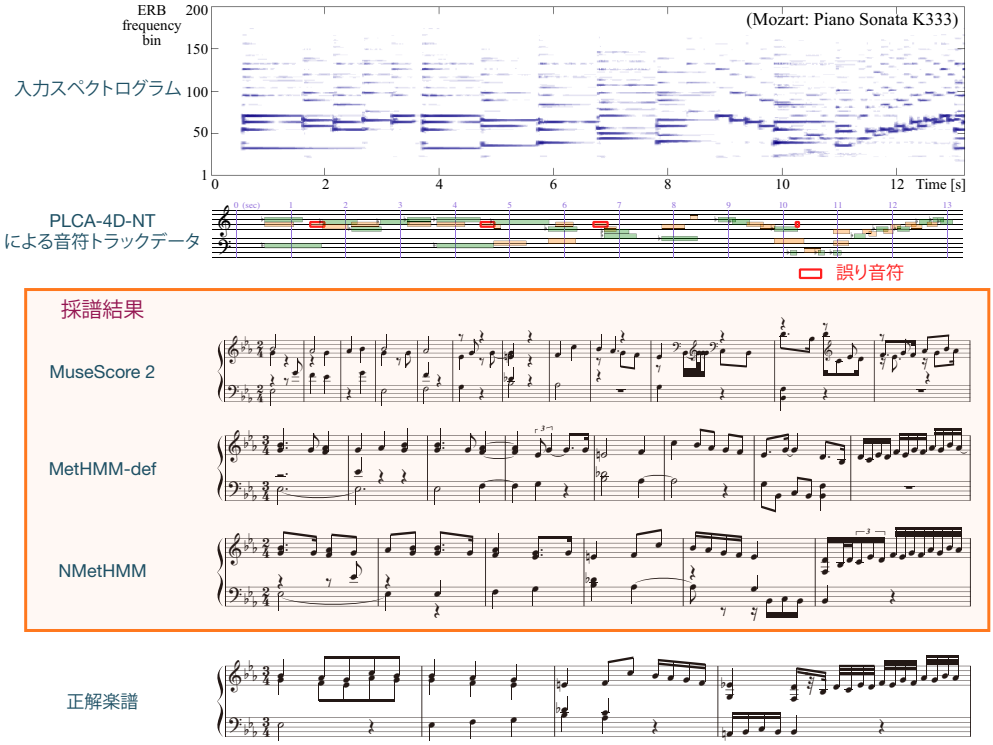

図12 採譜結果の例

3-6 まとめ

本研究では、従来別々に研究されてきた多重音検出とリズム量子化手法の最新手法を統合し、音響信号から楽譜を出力する多声音楽の採譜システムの構築を行い、系統的評価を行った。さらに、多重音検出とリズム量子化の両方において、従来手法の改良を行い、これが採譜の精度向上に有効であることを確かめた。また推定結果の楽譜に任意の誤りが含まれる場合にも適用可能な、評価尺度を開発した。システム構成の方法や評価の方法論は異なるアプローチを用いる研究においても適用可能なものであり、評価結果は今後の採譜研究において基準としての役割を果たすと期待される。

現状での採譜結果は、音響および音楽的に単純なケースでは、再現性と演奏可能性の上で比較的優れた結果が得られたものの、それ以外の場合は満足いく結果とは程遠い。今後の課題として、まずリズム量子化ステップにおいて音高情報を用いて、不自然なジャンプや演奏不可能な箇所を減らすことが考えられる。また現状のモデルでは、多重音検出によって生じた余分な音符を削減することはできるが、不足した音符を補うことや音高誤りを訂正することは不可能である。これを実現するためには、音響モデルと記号モデルのさらに実質的な統合が必要であると考えられる。また採譜結果の演奏可能性や音楽的観点からの主観評価を実施することも今後の課題である。

本節の内容およびその詳細は以下の国際会議論文に記載される予定である（採択済）。

・E. Nakamura, E. Benetos, K. Yoshii, and S. Dixon, “Towards Complete Polyphonic Music Transcription: Integrating Multi-Pitch Detection and Rhythm Quantization,” Proc. ICASSP, to appear, 2018.

4 確率的音型モデルの構築

4-1 背景

楽譜は音符からなる時系列であるが、そこには局所的な系列依存性の他に、句構造や階層構造、そして反復構造が存在する。これらの構造を取り込める楽譜モデルを作ることは自動編曲や採譜技術を含む幅広い応用にとって重要と考えられる。特に反復構造は、音楽に特有かつ広く見られる構造であり、予測性能の高いモデルを構成する上で特に大切だと考えられる。実際に我々の先行研究[9]では、楽譜のリズム音型に現れる反復構造を捉えたモデルを構築することで、採譜の性能が向上することが示された。その研究では、反復構造が音型の分布のスパースネスにより統計的に記述できることに着目したモデル定式化がなされた（図 13）。

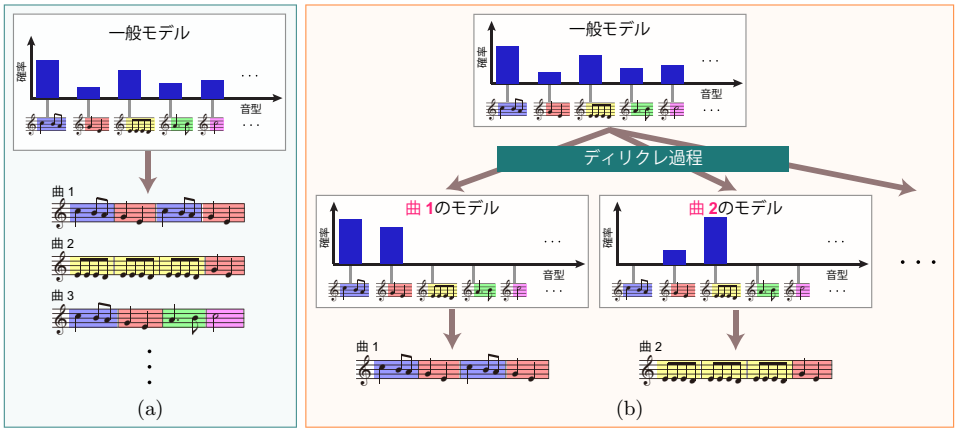

図13 一般モデルとディリクレ過程を用いた反復構造の記述

自動編曲や音響からの採譜にこの反復構造の統計的な記述法を適用するには、音高情報を含むモデル化が不可欠である。しかし、単純に音高とリズムの音型を考えたのでは、音型の数が組み合わせ爆発により巨大になり、計算量が実行不可能な大きさになってしまう。この問題の解決法として、本研究では音型の確率的記述を定式化する。この方法では、音型自体の記述に確率モデルを組み込むことで、音型の空間を直接扱うことを避けながら生成モデルを構成することができるため、計算量の問題を回避することが可能である。

4-2 提案手法

まず、リズムのみの場合を考える。リズムの楽譜モデルとして従来から拍節マルコフモデル[7]が知られているが、これは小節（あるいは音型の単位）内を考えると LR（left-to-right）型のマルコフモデルとして記述される。音型の確率的記述の鍵となるのは、拍節マルコフモデルを音型の集合と捉えることである。特に遷移確率行列がバイナリ行列として与えられるような拍節マルコフモデルは、ある特定の音型のみを生成するため、音型と一対一関係がある。一方で、一般の遷移確率行列の場合は、変形を含む音型の集合を定義していると解釈できる。よって拍節マルコフモデルの混合モデルを考えることで変形を含む音型の集合のモデルが記述できる。モデルパラメータは EM アルゴリズムにより学習可能である。

次に、音高も含むモデルを考える。この場合は、各ビート位置に対応して記号（音高）の出力確率を考えることにより拍節マルコフモデルの拡張として定式化される。この場合も、出力確率がバイナリベクトルで表される場合は、特定の音高パターンが表現されるが、一般の出力確率の場合は変形を含む音型の集合を表現可能となる（図 14）。

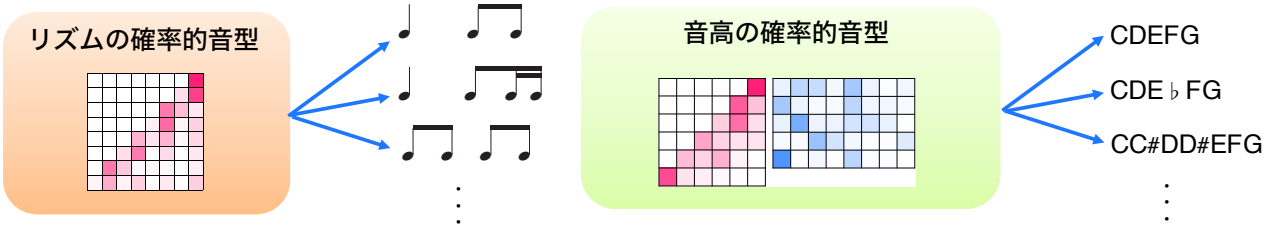

図14 確率的音型モデル（リズム音型と音高の音型）

以上の基本モデルの拡張により、反復構造を捉えた音型に基づく生成モデルが構成できる。まず全体のモデルをベイズ拡張する。この際、混合比は離散分布なので、事前分布として共役事前分布のディリクレ分布を用いるが、この集中度を小さくとることで、スパースネスが誘導され、結果として楽曲に用いられる音型の実質的な数が少なくなる。これにより反復構造が誘発されるモデルとなる。また混合比の代わりに、終状態を持つマルコフモデルを考えれば、曲の大局構造を反映したモデルとなる。これらを採譜に応用するため

には、音響モデルあるいは演奏モデルとの統合を行い、また編曲に応用するためには、楽譜の変形操作モデルとの統合を行うことで定式化ができる。

4-3 結果

ポピュラー音楽のメロディーデータを用いて、上記の確率的モデルの学習を行った結果を説明する。混合数が小さい場合のリズム音型の学習結果を図 15 に示す。この結果より特定の音価の使用頻度や連鎖の頻度などのリズムの特徴ごとに音型が自動でまとめられていることが分かる。またそれぞれの音型の集合は互いに音符分割などの変形操作でしばしば関係していることが見て取れる。混合数を増加させると、それぞれの音型の分布はよりスパースになり、それぞれの音型がより特定の音型を記述するようになることが確認された。これにより提案モデルの学習により変形を含めた確率的な音型の学習が可能であることが確認できる。

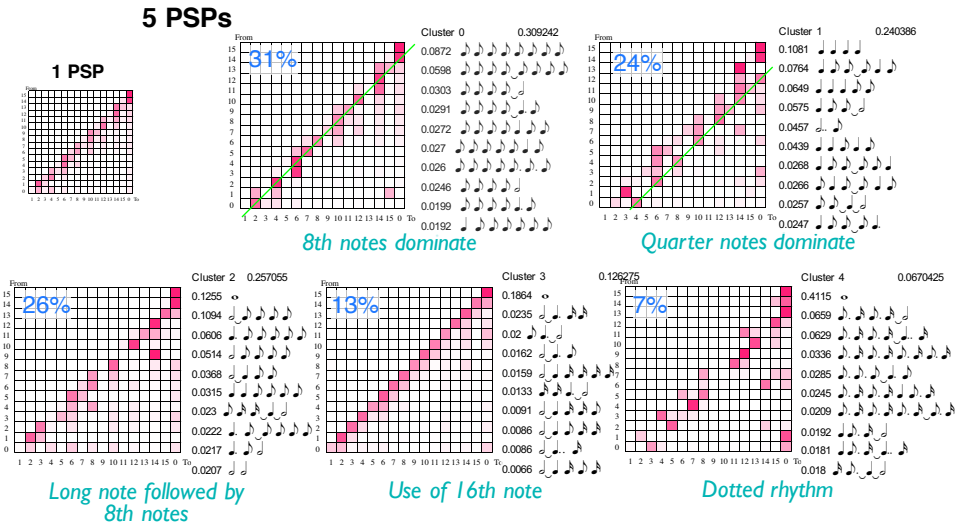

図15 確率的リズム音型モデルの学習結果

音高の音型の学習も同様の結果が得られている。この場合には、音符分割の他に音高シフトも変形パターンとして自動的に学習されることが確認できる。計算時間の詳しい測定は今後の課題であるが、混合数 100 程度でも学習可能であることは既に確認されている。また混合数を増加させることでパープレキシティーが低くなることも確認され、提案モデルが予測性能の向上に有効であることが示唆されている。

4-4 まとめ

現実的な計算量で学習・推論可能な音高を含む音型モデルの構築を行い、その学習動作を確認した。従来は計算量の問題で扱えなかった、採譜や編曲モデルにこの音型モデルを組み込むことにより、幅広い応用が期待される。現在メロディーの様式変換の編曲や歌声の採譜における有効性を検証中である。

今後の課題として、多声音楽への拡張がある。対位法的なメロディーが合わさった構造を持つ曲においては、各声部に提案モデルを適用して、その出力を合流する定式化が考えられる。一方で、ピアノのように和音を含む声部を持つ音楽に関しては、音高方向の依存性が重要であり、バスあるいはスカイラインに対する内声部の依存性を導入したモデルなどが考えられる。

以上の結果は、追加の結果と合わせて論文投稿を現在準備中である。

5 まとめと展望

以上、本研究で得られた成果を 3 つの主要トピックに分けて述べた。全体として、統計的手法に基づく自動編曲、多声音楽の自動採譜、そして音楽特有の構造を捉えた楽譜の生成モデルに関して進展があったと言える。今後、これらの成果を結合し、音楽スタイルの変換などの自動編曲システムや、さらに音響信号から編曲が行えるシステムの開発に応用することが考えられる。

本研究の定式化は、計算論的音楽分析や音楽分類などにも適用可能であり、今後さらに広い応用が見込まれる。音高を含む音型モデルは、楽曲に現れる特徴的な音型を学習できるのみならず、シンコペーションや非和声音など音楽のスタイルを特徴付ける特徴量としても機能すると考えられる。教師なし学習に基づく提案法を用いることにより、少量のデータしか手に入らない音楽や専門知識が発展していない音楽分野においても系統的な音楽分析が行える可能性がある。また自動採譜技術と融合させることにより、インターネット上に大量にある音楽音声データから直接、音楽的な情報解析が行える可能性もある。

本研究では主にマルコフモデルやその拡張を基にした統計モデル手法を扱ったが、近年、深層ニューラルネットを用いた音楽情報処理手法も発展してきている。深層学習では対象とする入力データと出力データの

関係を直接学習させる、End-to-End の手法が盛んに研究されている。一方で音楽では構造が複雑であり、また大量のデータが簡単に入手できないこともしばしばあり、楽譜や演奏の構造を反映したネットワークを考えることの必要性も意識されている。本研究の結果は、多声部構造や反復構造などといった音楽で普遍的に見られる構造のモデル定式化に関するものであり、深層学習の場合でもネットワークの設計指針などに示唆を与えるものと考えられる。

【参考文献】

[1] E. Benetos et al., “Automatic Music Transcription: Challenges and Future Directions,” J. Intelligent Information Systems, vol. 41, no. 3, pp. 407–434, 2013.

[2] E. Nakamura et al., “Rhythm Transcription of Polyphonic Piano Music Based on Merged-Output HMM for Multiple Voices,” IEEE/ACM Transactions on Audio, Speech and Language Processing, pp. 794–806, 2017.

[3] S.Young et al.,“Tree-Based State Tying for High Accuracy Acoustic Modelling,” Proc. Human Lang. Techno., pp. 307–312, 1994.

[4] D. Temperley, “A Unified Probabilistic Model for Polyphonic Music Analysis,” J. New Music Res., vol. 38, no. 1, pp. 3–18, 2009.

[5] E. Nakamura et al., “Merged-Output HMM for Piano Fingering of Both Hands,” in Proc. ISMIR, 2014, pp. 531–536.

[6] E. Benetos and T. Weyde, “An Efficient Temporally-Constrained Probabilistic Model for Multiple-Instrument Music Transcription,” in Proc. ISMIR, pp. 701–707, 2015.

[7] C. Raphael, “A Hybrid Hraphical Model for Rhythmic Parsing,” Artificial Intelligence, vol. 137, pp. 217–238, 2002.

[8] E. Vincent et al., “Adaptive Harmonic Spectral Decomposition for Multiple Pitch Estimation,” IEEE TASLP, vol. 18, no. 3, pp. 528–537, 2010.

[9] E. Nakamura et al., “Rhythm Transcription of MIDI Performances Based on Hierarchical Bayesian Modelling of Repetition and Modification of Musical Note Patterns,” Proc. EUSIPCO, pp. 1946–1950, 2016.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
Note Value Recognition for Piano Transcription Using Markov Random Fields	IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), Vol. 25, No. 7, pp. 1542-1554	2017 年 7 月
Towards Complete Polyphonic Music Transcription: Integrating Multi-Pitch Detection and Rhythm Quantization	IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)	2018 年 4 月（予定）