Vol.2019-MUS-125 No.1
2019/11/19

深層多重音検出を用いた音響信号から楽譜へのピアノ採譜

柴田 健太郎[1,a)]　中村 栄太[1,2,b)]　錦見 亮[1,c)]　吉井 和佳[1,d)]

概要：本稿では多重音響信号から楽譜を推定するピアノ自動採譜システムについて述べる．最近の研究で，多重音検出とリズム量子化手法の統合による自動採譜システムの可能性が示されたが，多重音検出の誤りによって採譜精度が頭打ちになってしまう問題があった．一方，深層学習技術が飛躍的に進歩しているため，本研究では現在最高精度の深層多重音検出手法を統合した新たな採譜システムを構築し，その効果を検証する．具体的には，多重音を含む音響信号から音符の音高，発音時刻，消音時刻，ベロシティーを推定する新たな畳み込み型のニューラルネットワークを提案する．また，和音を含む声部を扱うことができる声部分離手法を新たに提案する．評価実験により提案システムの採譜精度が従来手法を大幅に上回ることを示すと同時に，提案する声部推定によって楽譜の可読性が向上することを示す．

1. はじめに

自動採譜は音楽情報処理の分野において長年の未解決問題である [1–3]．自動採譜の最終目標は音楽音響信号を完全な楽譜へ変換することであり，得られた楽譜は音楽演奏や記号領域での音楽内容解析に役立つ．多声音楽の自動採譜はその問題の難しさから，多重音検出とリズム量子化の２つの問題に分けて主に研究されてきた．多重音検出は，音響信号を半音単位の音高と発音・消音時刻で表される音符列，およびベロシティー（音強）の系列である演奏 MIDI 系列（ピアノロール）へと変換する [4–9]．リズム量子化は，演奏 MIDI 系列を発音・消音時刻が拍単位で記述される量子化 MIDI 系列へと変換する [10–13]．

自動採譜に関するの研究が多く行われるなか，音響信号から完全な楽譜まで推定する研究の数は限られる [14–17]．文献 [14] は多重音検出とリズム量子化を多段処理の枠組みで統合したピアノ採譜手法を提案している．この手法は，当時最も性能が良い手法の 1 つであった確率的潜在成分分析（probabilistic latent component analysis; PLCA）を多重音検出に用いている．PLCA の出力は非常に多くの誤りを含むため，実用には程遠い採譜結果がしばしば得られた．一方，深層ニューラルネットネットワーク（Deep Neural Network; DNN）[7–9] を用いた多重音検出法が目覚ましい成果を挙げており，音響から楽譜への採譜における大幅な性能向上が期待できる．

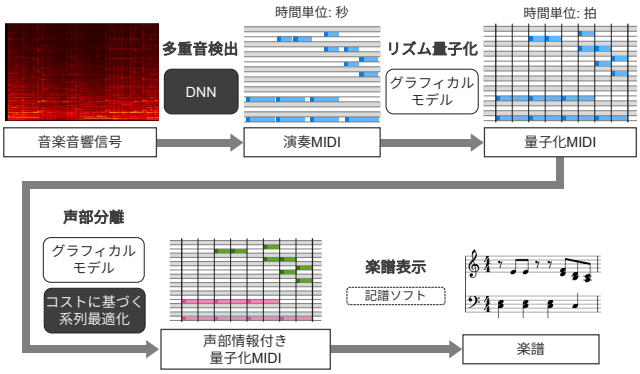

図 1　深層多重音検出を用いた音響信号から楽譜へのピアノ採譜手法

また，ピアノ採譜において声部分離は，読みやすい楽譜の生成に必要な，重要な問題の一つである [18–24]．ピアノ楽譜の声部構造は階層的である．まず，通常右手と左手に対応する二つのパートがある．さらにそれぞれのパートの中に複数の声部があり，それら声部は楽譜上では異なるレイヤーで表記される．既存手法の大半は各声部が単旋律だと仮定しているため，各声部が和音を含むことが多いピアノ楽譜の採譜には適していない．文献 [14] では記譜ソフトに実装されている声部分離手法を用いており，不要なタイや誤った場所に配置された休符によって出力楽譜の読みやすさが損なわれている．

本研究の目的は，最高精度のピアノ音響信号に対する自動採譜法のベンチマークを確立するとともに，現状での限界と残された主要な問題を示すことである．最新の多重音検出・リズム量子化・声部分離手法を統合することでピアノ採譜システムを構築する．現在フレーム単位の多重音推

1　京都大学 大学院情報学研究科
2　京都大学 白眉センター
a)　shibata@sap.ist.i.kyoto-u.ac.jp
b)　enakamura@sap.ist.i.kyoto-u.ac.jp
c)　nishikimi@sap.ist.i.kyoto-u.ac.jp
d)　yoshii@kuis.kyoto-u.ac.jp

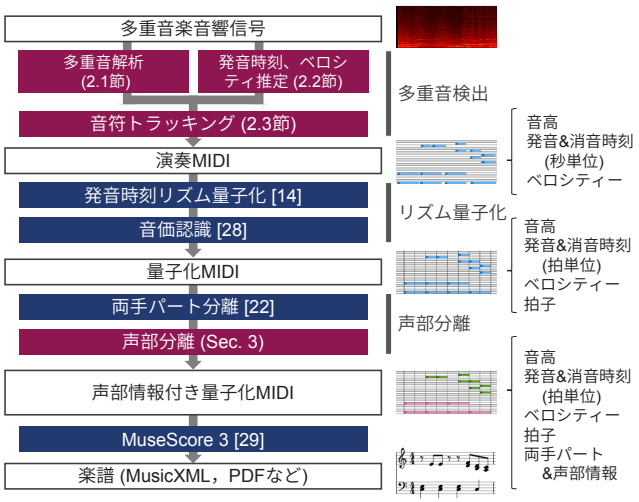

図 2　提案システムの構成

定で最高性能を達成している手法 [9,25] に基づき，新たな多重音検出法を提案する．また，和音を含む声部を扱えるコストに基づく声部分離法を提案する．MAPS データセット [26] を用いて，文献 [14,27] で提案された評価尺度に基づく系統的な評価実験を行う．また，提案する声部推定手法の有効性を確かめるための主観評価実験を行う．

提案する採譜システムの構成は文献 [14] で提案されたものの拡張であり図 2 に示す通りである．我々は新たな多重音検出法を提案するとともに（2 章），代表的な手法 [8] との比較も行う．リズム量子化ステップでは，拍節隠れマルコフモデル（hidden Markov model; HMM）を用いて拍単位の発音時刻を推定する（拍子が常に 4 / 4 拍子に固定されることを除き，文献 [14] の 4.1 章と同様）．その後, 拍単位の消音時刻をマルコフ確率場モデル [28] を用いて推定する．声部分離ステップでは，まず初めに両手パート分離手法 [22] を適用する．この手法は，一般的な MIDI 系列を入力として受け取り，右手と左手に対応する 2 つの MIDI 系列を出力する．その後，3 章で述べる声部分離の新手法を適用する．最終ステップでは，公開ソフト MuseScore 3 [29] を使用して，楽譜のグラフィカルな表現を得る．

2. 多重音検出

提案する多重音検出法（POVNet）は 2 つの DNN から構成される．1 つは音高解析用のネットワーク（PitchNet），もう 1 つは発音時刻とベロシティーの推定用のネットワーク（OnVelNet）である（図 3）．これらのネットワークは，文献 [9] で初めて多重音解析に用いられた，DeepLabv3+ [30] と呼ばれる画像領域分類用の畳み込みニューラルネットワーク（convolutional neural network; CNN）に基づく．各ネットワークを別々に学習し，それらの出力を音符トラッキングステップで統合することで演奏 MIDI 系列が得られる．両ネットワークは HCFP（harmonic combined frequency and periodicity）特徴量 $\mathbf{Z} \in \mathbb{R}_{+}^{2H\times F\times T}$ [9] を入力とする．

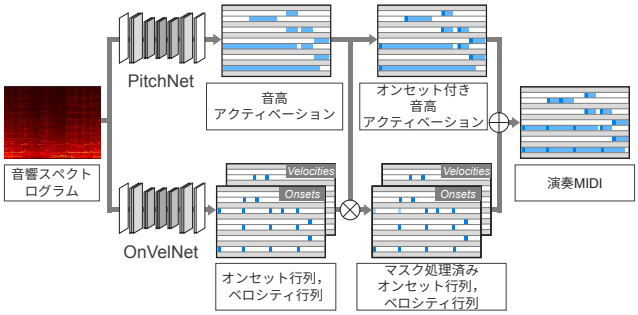

図 3　多重音検出ネットワーク（POVNet）の構成図．

ここで，H，F，T はそれぞれ調和音の数，周波数ビンの数，時間フレームの数である．本稿では文献 [9] と同様に $H = 6$，$F = 352 = 88 \times 4$ とした．

2.1 多重音解析ネットワーク（PitchNet）

HCFP 特徴量 $\mathbf{Z}$ を与えると，PitchNet は確率行列 $\mathbf{P}_{\mathrm{p}} \in [0,1]^{F\times T}$ を出力する．ここで，$\mathbf{P}_{\mathrm{p}}(f,t)$ はフレーム t における周波数 f のサリエンス（顕著さ）を表す．入力が $2H = 12$ チャネル，出力が 1 チャネルであることを除いて，PitchNet は文献 [9] で提案されたものと同じネットワークである．最終層でシグモイド関数を用いて $\mathbf{P}_{\mathrm{p}}$ が得られる．PitchNet は以下の損失関数を最小化するよう学習される．

$$\mathcal{L}_{\mathrm{p}} = -\frac{1}{F\times T}\sum_{f=1}^{F}\sum_{t=1}^{T}\hat{\mathbf{P}}_{\mathrm{p}}(f,t)\log \mathbf{P}_{\mathrm{p}}(f,t) \qquad (1)$$

ここで，$\hat{\mathbf{P}}_{\mathrm{p}} \in \{0,1\}^{F\times T}$ は正解 MIDI データ（サスティンペダルイベントを含む）をアップサンプルして得られたバイナリ行列である．最後に，$\mathbf{P}_{\mathrm{p}}$ を 2 値化し，ダウンサンプルすることで音高アクティベーション行列 $\mathbf{D}_{\mathrm{p}} \in \{0,1\}^{M\times T}$ が得られる．ここで，$\mathbf{D}_{\mathrm{p}}(m,t)$ はフレーム t における半音単位の音高 m の音の有無を表し，$M = 88$ はピアノの鍵盤の個数を表す．

2.2 発音時刻・ベロシティー推定ネットワーク（OnVelNet）

HCFP 特徴量 $\mathbf{Z}$ を与えると，OnVelNet はオンセット確率行列 $\mathbf{P}_{\mathrm{o}} \in [0,1]^{F\times T}$ と音強行列 $\mathbf{P}_{\mathrm{v}} \in [0,1]^{F\times T}$ を出力する．ここで，$\mathbf{P}_{\mathrm{o}}(f,t)$ と $\mathbf{P}_{\mathrm{v}}(f,t)$ はそれぞれ，フレーム t，周波数 f におけるオンセットサリエンスと音強を表す．また，音強は MIDI 規格でベロシティーとして定められた 0 から 127 まで整数値をリスケールした 0 から 1 までの値をとる．出力がオンセット行列用と音強行列用の 2 チャネルであることを除いて，OnVelNet は PitchNet と同じ構造である．最終層では，シグモイド関数を用いて $\mathbf{P}_{\mathrm{o}}$ が，$[0,1]$ 区間のクリップ関数を用いて $\mathbf{P}_{\mathrm{v}}$ がそれぞれ得られる．このネットワークは以下で定義されるバイナリクロスエントロピー損失と $\mathcal{L}_{\mathrm{o}}$ と平均二乗誤差損失 $\mathcal{L}_{\mathrm{v}}$ の重み付き和を最小化するよう学習される．

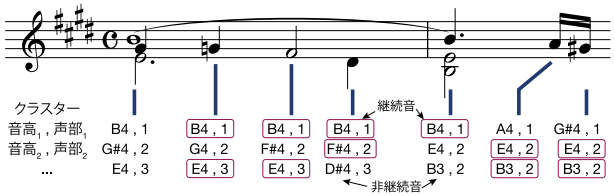

図 4 声部配置の例

$$\mathcal{L}_{\mathrm{o}} = -\frac{1}{F \times T} \sum_{f=1}^{F} \sum_{t=1}^{T} \hat{\mathbf{P}}_{\mathrm{o}}(f,t) \log \mathbf{P}_{\mathrm{o}}(f,t), \tag{2}$$

$$\mathcal{L}_{\mathrm{v}} = \frac{1}{F \times T} \sum_{f=1}^{F} \sum_{t=1}^{T} \hat{\mathbf{P}}_{\mathrm{o}}(f,t) \left(\hat{\mathbf{P}}_{\mathrm{v}}(f,t) - \mathbf{P}_{\mathrm{v}}(f,t) \right)^2, \tag{3}$$

$$\mathcal{L}_{\mathrm{ov}} = w_{\mathrm{o}} \mathcal{L}_{\mathrm{o}} + w_{\mathrm{v}} \mathcal{L}_{\mathrm{v}}, \tag{4}$$

ここで，$\hat{\mathbf{P}}_{\mathrm{o}} \in \{0,1\}^{F \times T}$ は正解音符のオンセットの有無を表すバイナリ行列，$\hat{\mathbf{P}}_{\mathrm{v}} \in [0,1]^{F \times T}$ は正解音符のオンセットにおける音強を表す実数行列である．$\hat{\mathbf{P}}_{\mathrm{o}}$ と $\hat{\mathbf{P}}_{\mathrm{v}}$ は 2.1 節における $\hat{\mathbf{P}}_{\mathrm{p}}$ と同様の方法で正解 MIDI から得られる．発音時刻のわずかな揺らぎを許容するため，$\hat{\mathbf{P}}_{\mathrm{o}}(f,t) = 1$ ならば $\hat{\mathbf{P}}_{\mathrm{o}}(f,t \pm 1) = 1$ とした．文献 [8] と同様に $\mathcal{L}_{\mathrm{v}}$ は正解のオンセットでマスクされた後で計算される．

最後に，$\mathbf{P}_{\mathrm{o}}$ を二値化しダウンサンプルすることでオンセット行列 $\mathbf{D}_{\mathrm{o}} \in \{0,1\}^{M \times T}$ が得られる．$\mathbf{D}_{\mathrm{o}}$ において時間方向に連続した 1 の要素は中央の要素を除いて 0 に設定する．ベロシティー行列 $\mathbf{D}_{\mathrm{v}} \in \{0,\ldots,127\}^{M \times T}$ は，$\mathbf{P}_{\mathrm{v}}$ に対してスケーリング，整数化した後ダウンサンプルすることで得られる．

2.3 音符トラッキング

演奏 MIDI 系列 $\mathbf{D}_{\mathrm{n}}$ は音符の音高，発音時刻，消音時刻とベロシティーで構成され，上述の音高アクティベーション行列 $\mathbf{D}_{\mathrm{p}}$, オンセット行列 $\mathbf{D}_{\mathrm{o}}$, ベロシティー行列 $\mathbf{D}_{\mathrm{v}}$ から求められる．$\mathbf{D}_{\mathrm{n}}$ のオンセットは $\mathbf{D}_{\mathrm{p}}(f,t) = 1$ かつ $\mathbf{D}_{\mathrm{p}}(f,t-1) = 0$ の条件を満たす TF ビンを列挙することで得られる．$\mathbf{D}_{\mathrm{o}}$ は同音連打オンセットを求めるためだけに用いられるが，音高行列 $\mathbf{D}_{\mathrm{n}}$ との不整合が生じない様に，$\mathbf{D}_{\mathrm{p}}(f,t) = 0$ となる TF ビンでは $\mathbf{D}_{\mathrm{o}}(f,t)$ は 0 となるようにマスクをかける．最終的に $\mathbf{D}_{\mathrm{n}}$ は以下の規則によって生成される．

- $\mathbf{D}_{\mathrm{p}}$ によって得られるオンセットとマスクされた $\mathbf{D}_{\mathrm{o}}$ から得られるオンセットが 50 ms 以内に入っている場合，先行するオンセットのみを採用する．
- 最終的な音符継続長が 30 ms 以内の音符は除去する．

3. ピアノ採譜のための声部分離

我々の提案する声部分離手法は，声部の適切な構造と，入力 MIDI 系列に対する声部の適合度を表現するコスト関数に基づいた系列最適化手法である．入力 MIDI 系列は右手左手パートに分離済みとし，本手法では右手左手パートに対して独立に声部分離を行う．系列最適化では発音時刻が同じである音符の組を単位として扱う．発音時刻が同じである音符の組に対してクラスターを構成し，それよりも発音時間が前で音がオーバーラップする音符（継続音と呼ぶ）もそのクラスターに加える．クラスター内の音符に対する声部を整数のラベル $1, 2, \ldots, V_{\max}$ で記述する．ここで，声部数の上限値 $V_{\max}$ は任意に設定できる変数である．各クラスター C_k に対して，クラスター内の音符 $n \in C_k$ の声部ラベル $S_k = (s_n)$ の組を声部配置と呼ぶ．声部分離の探索空間は全てのクラスターに対する全ての可能な声部配置の集合である．

コスト関数は，以下で定義される垂直コストと水平コストの和で構成される．垂直コスト $V(S_k)$ は，あるクラスターに対する声部配置の適切さを記述し，以下の 4 つの因子の和で定義される．

- 各音符 $n \in C_k$ に s_n を与える（不要な声部にペナルティを与える）．
- 声部の順番と音高の順番が逆になっている音符のペアそれぞれに λ_2 を与える（声部の交差にペナルティを与える）．
- 声部ラベルが同じで消音時刻が異なる音符のペアそれぞれに λ_3 を与える．
- 同じ声部ラベルを持つ継続音と継続音でない音符のペアそれぞれに λ_4 を与える．

水平コスト $H(S_{k-1}, S_k)$ は隣接するクラスターの声部配置の適切なつながりを記述し，以下の 3 つの因子の和で定義される．

- 声部ラベルが一貫していない継続音に λ_5 を与える．
- 同じ声部ラベルを持ちながら時間の開きがある隣接する音符のペアそれぞれに λ_6 を与える．(休符にペナルティを与える).
- 同じ声部ラベルを持ちながら時間的オーバーラップがある隣接する音符のペアそれぞれに λ_7 を与える．

系列最適化は Viterbi アルゴリズムによって解くことができる．声部内の和音の消音時刻は一致せねばならず，かつその声部の次の音符の発音時刻と同じかそれより前でないといけないという声部の制約を満たすために，声部推定後に，消音時刻は得られた声部にしたがって補正される．上記のコスト関数の重み付けはいくつかの試行により $(\lambda_2, \ldots, \lambda_7) = (3, 1, 1, 5, 0.2, 1)$ としたが，系統的なパラメータ最適化の余地が残っている．

4. 評価

4.1 実験設定

多重音検出の精度を評価するために MAPS データベース [26] の “ENSTDkCl” と “ENSTDkAm” のラベルが付い

手法	フレーム単位			ノート単位		
	$\mathcal{P}_\mathrm{f}$	$\mathcal{R}_\mathrm{f}$	$\mathcal{F}_\mathrm{f}$	$\mathcal{P}_\mathrm{n}$	$\mathcal{R}_\mathrm{n}$	$\mathcal{F}_\mathrm{n}$
PLCA [14]	—	—	—	77.9	68.9	72.8
Onsets and Frames [33]	**92.9**	78.5	84.9	87.5	**85.6**	86.4
DeepLabv3+ [9]	87.5	**86.3**	86.7	—	—	—
PitchNet のみ	89.3	84.4	86.6	91.1	68.4	77.5
POVNet（提案手法）	89.3	85.7	**87.3**	**89.7**	84.1	**86.7**

表 1 多重音検出精度 (%)

た 60 曲を用いた．また，提案システムの最終的な出力である楽譜の採譜精度を評価するためには “ENSTDkCl” のラベルが付いた 30 曲を用いた．従来手法 [14] と同様に，楽譜の評価にはそれぞれの録音の最初の 30 s を評価の対象とした．PitchNet と OnVelNet の学習には MAPS データセットのうち上記の 60 曲を除く全ての曲を用いた．ネットワークのパラメータは最適化手法 RAdam [31] を用いて更新し，学習率の初期値は 0.001 とした．OnVelNet の損失関数の重みはそれぞれ $w_\mathrm{o} = 0.9$， $w_\mathrm{v} = 0.1$ とした．

2 章の多重音検出の性能評価には文献 [32] で導入されているフレーム単位及び音符単位の評価尺度を用いた．フレーム単位の尺度では 10 ms の時間分解能で適合率 $\mathcal{P}_\mathrm{f}$，再現率 $\mathcal{R}_\mathrm{f}$，F 値 $\mathcal{F}_\mathrm{f}$ を定義する．音符単位の尺度では音符の発音時刻に基づいて評価され，推定音符の発音時刻が正解音符の発音時刻の ±50 ms 以内に入っていれば正解とみなされる．この基準に基づき音符単位の適合率 $\mathcal{P}_\mathrm{n}$，再現率 $\mathcal{R}_\mathrm{n}$，F 値 $\mathcal{F}_\mathrm{n}$ を定義する．採譜システム全体の性能評価には文献 [14] にて導入された編集距離に基づく評価尺度と文献 [27] にて導入された評価尺度 MV2H を用いた．前者では以下の誤り率を定義する：音高誤り率 E_p，不足音符率 E_m，余分音符率 E_e，発音時刻誤り率 E_on，消音時刻誤り率 E_off とこれら 5 つの尺度の平均の総合誤り率 E_all である．評価尺度 MV2H では，以下の精度を定義する：多重音検出 $\mathcal{F}_\mathrm{p}$，声部分離 $\mathcal{F}_\mathrm{voi}$，拍節配位 $\mathcal{F}_\mathrm{met}$, 音価推定 $\mathcal{F}_\mathrm{val}$, 和声解析 $\mathcal{F}_\mathrm{harm}$ とこれらの平均 $\mathcal{F}_\mathrm{MV2H}$．ここで，文献 [27] では $\mathcal{F}_\mathrm{harm}$ がコードラベル推定精度と調推定精度の重み付き和として定義されているが，提案システムではコードラベルは出力しないので，調推定精度を意味する．

4.2 結果

表 1 に多重音検出手法の精度を示す．提案手法（POVNet）との比較として，文献 [14] の PLCA に基づく手法，文献 [33] の DNN に基づく代表的な手法（MAESTRO データベースで学習した Onsets and Frames），および文献 [9] の手法（DeepLabv3+）による結果を示す．さらに，OnVelNet を用いない PitchNet のみによる結果も示すが，これは DeepLabv3+ とほぼ等価である．表より，我々の提案する POVNet がフレーム単位と音符単位両方の F 値において比較手法を上回ることが見て取れる．POVNet と Onsets and Frames は $\mathcal{F}_n$ において同等な精度となっているが，PLCA に基づく手法に対しては大きな差が見られる．PitchNet のみと POVNet の $\mathcal{R}_\mathrm{n}$ の間に大きな差があることから，OnVelNet によって同音連打を扱えるようになった効果が見て取れる．

表 2 に採譜システム全体の評価の結果を示す．深層音高検出によって採譜精度が大幅に改善していることが見て取れる．編集距離に基づく評価尺度上では，多重音検出機構に PONVnet を用いたシステムが Onsets and Frames [33] を用いたシステムより総じて精度がよい．POVNet+RQ と POVNet+RQ+VS の結果を見比べることで，提案した声部推定法がわずかながら編集距離基準の誤り率を低下させることが分かる．声部推定手法を用いない POVNet+RQ が $\mathcal{F}_\mathrm{voi}$ において PLCA+RQ+VS 上回っていることは注目に値する．これに関しては，評価尺度 $\mathcal{F}_\mathrm{voi}$ では単旋律の声部のみを仮定しているため，ピアノ楽曲のように和音を含む声部分離の評価には適していない可能性も考えられる．

PLCA+RQ+VS の平均誤り率が Nakamura *et al.* [14] よりも悪い理由として，後者は余分な音符を省く機構を持つ Noisy 拍節 HMM を用いていることが考えられる．この余分な音を取り除く機構を現在のシステムに導入することも考えられるが，深層音高検出を用いたモデルでは適合率 $\mathcal{P}_\mathrm{n}$ が高いため，効果は限定的だと思われる．表 2 の一番下の行はリズム量子化と声部分離を正解演奏 MIDI に対して適用した結果を示している．この結果から，多重音検出の精度は，音高推定に関連する尺度のみならず，リズム採譜に関する尺度にも依然として大きな影響を与えることが見て取れる．

推定楽譜の一例を図 5 に示す．深層音高検出を用いた手法によって得られた楽譜は従来手法 [14] による楽譜より大幅に改善していることは明らかである．POVNet+RQ と POVNet+RQ+VS の結果を比較することで，声部分離手法が余分なタイや休符を効果的に減らしていることが見て取れる．また，全ての推定結果において図 5 の青枠で示す通り，後打音は正しく認識されておらず，これは我々のリズム量子化手法の現在の限界を示している．

さらに，声部推定が楽譜の可読性に与える影響を検証するために，主観評価実験を行った．POVNet+RQ+VS と POVNet+RQ による推定楽譜をどちらによるものか隠して２枚同時に被験者に見せ，「演奏するにあたっての楽譜の読みやすさ」が高い方，もしくは「どちらとも言えない」の選択肢を選んでもらった．実験に用いたプラットフォームはリンク先 [34] から確認できるようになっている．実験には，MAPS データセットのうち “ENSTDkCl” のラベルが付いた 30 曲を用いた．実験にはピアノ楽譜を読める 21 人が参加し，そのうち約 60% 人が 10 年以上のピアノ演奏経験者であった．得られた 630 の回答のうち，POVNet+RQ+VS が 496 回（78.7%），POVNet+RQ が 54 回（8.57%），「どちらとも言えない」が 80 回（12.7%）選ばれた．この結果は，提案した声部推定法により楽譜の可読性が向上することを

手法	E_{p}	E_{m}	E_{e}	E_{on}	E_{off}	E_{all}	$\mathcal{F}_{\mathrm{p}}$	$\mathcal{F}_{\mathrm{voi}}$	$\mathcal{F}_{\mathrm{met}}$	$\mathcal{F}_{\mathrm{val}}$	$\mathcal{F}_{\mathrm{harm}}$	$\mathcal{F}_{\mathrm{MV2H}}$
Nakamura ら [14]	2.92	28.1	11.4	19.4	41.8	20.7	68.1	45.8	31.1	80.5	57.2	56.5
PLCA [14]+RQ+VS	3.29	23.5	13.4	25.1	41.8	21.4	67.2	46.7	32.5	81.3	57.1	57.0
Onsets and Frames [33]+RQ+VS	0.96	8.23	7.36	**12.8**	28.1	11.5	82.8	46.7	41.6	86.6	**74.6**	66.4
POVNet+RQ	**0.74**	8.19	5.83	12.9	28.6	11.3	83.9	**48.2**	**43.6**	86.5	71.1	**66.7**
POVNet+RQ+VS（提案手法）	**0.74**	**8.11**	**5.75**	**12.8**	**27.9**	**11.0**	**84.3**	46.1	43.4	**87.3**	71.1	66.4
Ground truth MIDI+RQ+VS	0.45	2.60	2.46	5.36	21.3	6.44	90.4	49.2	47.4	86.7	77.8	70.3

表 2 MAPS-ENSTDkCl データセットに対する採譜の誤り率（%）と精度（%）．‘+RQ’ はリズム量子化法を適用したことを表し，‘+VS’ は我々が提案した声部分離法を適用したことを表す．‘+VS’ がない場合，声部分離には MuseScore 3 を用いたことを表す．

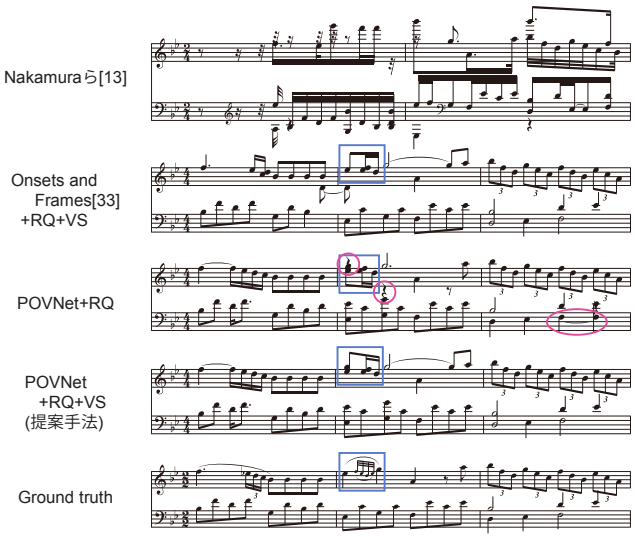

図 5 採譜結果の例 (MAPS-ENSTDKCl データセット，Mozart: Piano Sonata K. 333).

示している．

4.3 考察

実験により，我々の提案するピアノ採譜システムの楽譜推定精度は，従来の最高性能のシステムを大幅に上回ることを示した．デモページ[*1]に掲載された例から見て取れる様に，提案システムによって推定される楽譜はピアノ演奏のための楽譜として実用的なレベルに達していることも多い．しかし，まだ我々の提案するシステムには限界も多く，改善の余地が残されている．表 2 から見て取れる様に，多重音検出の改善によって最終的な楽譜の推定精度は更に向上する．同様にリズム量子化，特に消音時刻推定精度 E_{off} は正解 MIDI からの採譜結果でさえ 20% を越えており，改善により推定楽譜の品質は大きく向上すると思われる．この問題への解決策として，深層学習の技術を消音時刻推定に適用することと，声部配置と消音時刻を同時推定することが見込みがあると考えられる．さらに，消音時刻をより正しく推定するにはサステインペダルイベントの検出も重要である．

トリル，アルペジオ，前打音，後打音，グリッサンドといった装飾音はこれまで自動採譜の研究において扱われることは稀であったが，リズム量子化の精度を向上させるためには明示的なモデル化が不可欠である．完璧な楽譜の推定を目指す上で，ペダルイベント，強弱記号，スラー，アーティキュレーションや運指番号の推定も重要である．本研究ではクラッシック音楽のみで構成される MAPS データセットのみを用いたが，ジャズやポピュラー音楽を用いてモデルの学習及び検証を行うことも今後の課題である．

[*1] http://sap.ist.i.kyoto-u.ac.jp/members/shibata/sigmus125/

5. 結論

本稿では，最新の多重音検出，リズム量子化手法，および声部分離手法を統合することで，現在最高精度のピアノ音響信号に対する自動採譜システムを提案した．評価実験により，提案システムが既存システムの採譜精度を大きく上回ることを確認すると同時に，現在の自動採譜の限界と今後考えられるの研究の方向性を論じた．提案した手法により，実用レベルのピアノ自動採譜の実現が期待される．今後の研究でリズム量子化と声部推定に深層学習の技術を適用することで，さらなる精度向上を目指す予定である．また，今回提案したシステムをウェブサービスとして公開し，YouTube 等のピアノ演奏動画や任意のピアノ音響信号から楽譜を推定して提供する枠組みを構築する予定である．

6. 謝辞

本研究の一部は，JSPS 科研費 No. 16H01744，No. 19H04137，19K20340 および JST ACCEL No. JPMJAC1602 の支援を受けた．

参考文献

[1] Klapuri, A. and Davy, M.: *Signal Processing Methods for Music Transcription*, Springer Science & Business Media (2006).
[2] Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H. and Klapuri, A.: Automatic Music Transcription: Challenges and Future Directions, *Journal of Intelligent Information Systems*, Vol. 41, No. 3, pp. 407–434 (2013).
[3] Benetos, E., Dixon, S., Duan, Z. and Ewert, S.: Automatic Music Transcription: An Overview, *IEEE Signal Processing Magazine*, Vol. 36, No. 1, pp. 20–30 (2018).
[4] Benetos, E. and Weyde, T.: An efficient temporally-

constrained probabilistic model for multiple-instrument music transcription, *Proc. ISMIR*, pp. 701–707 (2015).

[5] Cheng, T., Mauch, M., Benetos, E., Dixon, S. et al.: An attack/decay model for piano transcription, *Proc. ISMIR*, pp. 584–590 (2016).

[6] Sigtia, S., Benetos, E. and Dixon, S.: An end-to-end neural network for polyphonic piano music transcription, *IEEE/ACM Transactions on Audio, Speech and Language Processing*, Vol. 24, No. 5, pp. 927–939 (2016).

[7] Bittner, R. M., McFee, B., Salamon, J., Li, P. and Bello, J. P.: Deep salience representations for F0 estimation in polyphonic music, *Proc. ISMIR*, pp. 63–70 (2017).

[8] Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J., Oore, S. and Eck, D.: Onsets and Frames: Dual-objective piano transcription, *Proc. ISMIR*, pp. 50–57 (2018).

[9] Wu, Y.-T., Chen, B. and Su, L.: Polyphonic Music Transcription with Semantic Segmentation, *Proc. ICASSP*, IEEE, pp. 166–170 (2019).

[10] Cemgil, A. T., Desain, P. and Kappen, B.: Rhythm Quantization for Transcription, *Comp. Mus. J.*, Vol. 24, No. 2, pp. 60–76 (2000).

[11] Raphael, C.: A Hybrid Graphical Model for Rhythmic Parsing, *Artificial Intelligence*, Vol. 137, pp. 217–238 (2002).

[12] Hamanaka, M., Goto, M., Asoh, H. and Otsu, N.: A Learning-Based Quantization: Unsupervised Estimation of the Model Parameters, *Proc. ICMC*, pp. 369–372 (2003).

[13] Nakamura, E., Yoshii, K. and Sagayama, S.: Rhythm transcription of polyphonic piano music based on merged-output HMM for multiple voices, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, No. 4, pp. 794–806 (2017).

[14] Nakamura, E., Benetos, E., Yoshii, K. and Dixon, S.: Towards complete polyphonic music transcription: Integrating multi-pitch detection and rhythm quantization, *Proc. ICASSP*, IEEE, pp. 101–105 (2018).

[15] Carvalho, R. G. C. and Smaragdis, P.: Towards End-to-End Polyphonic Music Transcription: Transforming Music Audio Directly to a Score, *Proc. WASPAA*, pp. 151–155 (2017).

[16] Román, M. A., Pertusa, A. and Calvo-Zaragoza, J.: An End-to-end Framework for Audio-to-Score Music Transcription on Monophonic Excerpts., *Proc. ISMIR*, pp. 34–41 (2018).

[17] Nishikimi, R., Nakamura, E., Fukayama, S., Goto, M. and Yoshii, K.: Automatic Singing Transcription Based on Encoder-Decoder Recurrent Neural Networks with a Weakly-Supervised Attention Mechanism, *Proc. ICASSP*, pp. 161–165 (2019).

[18] Chew, E. and Wu, X.: Separating voices in polyphonic music: A contig mapping approach, *Proc. CMMR*, pp. 1–20 (2004).

[19] Cambouropoulos, E.: Voice and stream: Perceptual and computational modeling of voice separation, *Music Perception: An Interdisciplinary Journal*, Vol. 26, No. 1, pp. 75–94 (2008).

[20] Temperley, D.: A unified probabilistic model for polyphonic music analysis, *Journal of New Music Research*, Vol. 38, No. 1, pp. 3–18 (2009).

[21] Duane, B. and Pardo, B.: Streaming from MIDI Using Constraint Satisfaction Optimization and Sequence Alignment, *Proc. ICMC*, pp. 1–8 (2009).

[22] Nakamura, E., Ono, N. and Sagayama, S.: Merged-Output HMM for Piano Fingering of Both Hands, *Proc. ISMIR*, pp. 531–536 (2014).

[23] McLeod, A. and Steedman, M.: HMM-based voice separation of MIDI performance, *Journal of New Music Research*, Vol. 45, No. 1, pp. 17–26 (2016).

[24] de Valk, R. and Weyde, T.: Deep neural networks with voice entry estimation heuristics for voice separation in symbolic music representations, *Proc. ISMIR*, pp. 281–288 (2018).

[25] 中村栄太，Benetos, E.，吉井和佳，Dixon, S.：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，*Proc. SIGMUS*, Vol. 2018-MUS-120, No. 19, pp. 1–6 (2018).

[26] Emiya, V., Badeau, R. and David, B.: Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 6, pp. 1643–1654 (2009).

[27] McLeod, A. and Steedman, M.: Evaluating Automatic Polyphonic Music Transcription., *Proc. ISMIR*, pp. 42–49 (2018).

[28] Nakamura, E., Yoshii, K. and Dixon, S.: Note Value Recognition for Piano Transcription Using Markov Random Fields, *IEEE/ACM Transactions on Audio, Speech and Language Processing*, Vol. 25, No. 9, pp. 1542–1554 (2017).

[29] : MuseScore3, `https://musescore.org/en`.

[30] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, *Proc. ECCV*, pp. 801–818 (2018).

[31] Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J. and Han, J.: On the Variance of the Adaptive Learning Rate and Beyond, *arXiv preprint arXiv:1908.03265* (2019).

[32] Bay, M., Ehmann, A. F. and Downie, J. S.: Evaluation of Multiple-F0 Estimation and Tracking Systems, *Proc. ISMIR*, pp. 315–320 (2009).

[33] Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.-Z. A., Dieleman, S., Elsen, E., Engel, J. and Eck, D.: Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset, *Proc. ICLR*, (online), available from ⟨https://openreview.net/forum?id=r1lYRjC9F7⟩ (2019).

[34] : 主観実験プラットフォーム，`https://forms.gle/iTnkNckGwBN3CFkV9`.