Vol.2021-MUS-132 No.18
2021/9/17

ピアノ採譜のための音価推定と声部分離のマルチタスク学習

平松 祐紀 [1,a)]　中村 栄太[1,2,b)]　吉井 和佳[1,3,c)]

概要：本稿では，ピアノ採譜のために音価と声部を同時に推定する手法について述べる．近年，深層学習の発展により，音高と発音時刻の推定精度は飛躍的に向上した．一方で，楽譜を書き起こすために必要な音価と声部の推定は，依然として難しい課題である．これまでの研究で，次のことがわかっている：(1) 音価の推定には，音高と発音時刻の情報は有用だが，演奏された音の長さはそこまで有用ではない．(2) 音価と声部には相互関係がある．そこで，本研究では事前に推定された音高と発音時刻の情報から，音価と声部を同時に推定する双方向長短期記憶 (BiLSTM) ネットワークを提案する．また，入力に含まれるテンポ誤りや音符の挿入・削除誤りに対する頑健性を向上するため，データ拡張を行う．実験により，提案手法の有効性を確認する．

1. はじめに

自動ピアノ採譜の最終的な目標は，ピアノ演奏の音響信号を人間が読める楽譜に変換することであり，これは音楽解析や演奏補助に有用な技術である [1]．ピアノ曲はポリフォニックであり，声部と呼ばれる音の流れが複数存在するため，ピアノ採譜は難しい問題である．これまでのピアノ採譜に関する研究では，楽譜ではなくピアノロールを推定するものが多い [2–7]．つまり，推定結果が音高については量子化されているが，リズムについては量子化されていない．近年，フレーム単位で音高を推定する多重音検出 [8–10] や，発音時刻の量子化 [11, 12] については著しい改善がみられているが，可読性の高い楽譜を書き起こすために必要な音価と声部ラベルの推定 [13] は，依然として難しい問題である．ここで，音価は楽譜上での各音符の継続時間を表し，声部ラベルは各音符が属する楽譜の上段または下段（右手パートまたは左手パート）の声部を表す（図 1）．

これまでに，声部ラベルのないピアノ楽譜に対して，声部ラベルを推定する手法が提案されている [14–17]．各声部はモノフォニックであり，音価が高精度で推定されているという仮定のもとでは，声部ラベルは正確に推定できる [18, 19]．しかしながら，実際には，各声部は同時発音の複数の音符（和音）からなるホモフォニックな構造を持ち，

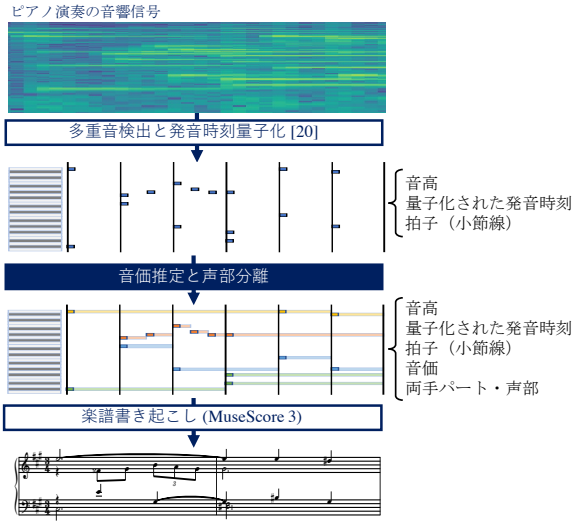

図 1　提案するピアノ採譜手法

音価の正確な推定は依然として難しい問題である．このような現実的な状況に対応するため，最先端のピアノ採譜手法では，声部ラベルの推定にルールベースの損失関数を使用しており，精度は限られている [20]．これを改良するためには，現代の深層学習に基づく統計的なアプローチが有用である．

音価推定の研究は比較的少なく [4, 21]，依然として難しい問題である [20]．中村ら [21] は，詳細な統計的分析を行い，次のことを明らかにした：(1) 音価を推定するとき，音高と発音時刻の情報は有用であるが，演奏された音の長さはそこまで有用ではない．(2) 音価と声部は相互関係を持つ．2 つ目の主張は，次の事実に基づくものである：一つの声部の中では，ある音符の消音時刻は次の音符の発音時刻に一致することが多い．すなわち，音符間に休符が挿入

[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
[3] 科学技術振興機構 戦略的創造研究推進事業（さきがけ）
[a)] hiramatsu@sap.ist.i.kyoto-u.ac.jp
[b)] enakamura@sap.ist.i.kyoto-u.ac.jp
[c)] yoshii@i.kyoto-u.ac.jp

されることは少ない [4]．よって，音価と声部を同時に推定することで，両問題の精度を改善できると考えられる．

本稿では，音響から楽譜へのピアノ採譜のために，事前に推定された音高と発音時刻の列から音価と声部ラベルを同時に推定する深層ニューラルネットワーク（DNN）を提案する．具体的には，双方向長短期記憶（BiLSTM）ネットワークをマルチタスク学習の枠組みに従って学習する．また，推定精度に影響を与えると考えられる，データ表現，ネットワークアーキテクチャ，推定結果の後処理，データ拡張を検討する．評価実験では，クラシックとポピュラー音楽のデータセットを使い，同時推定の有効性を検証する．

本研究の主な貢献は，深層学習に基づく音価と声部の同時推定手法を提案することである．最新のピアノ採譜手法 [20] で使われている，多重音検出手法と発音時刻量子化手法と統合することで，音響から楽譜へのピアノ採譜において，最先端の精度を実現した．もう一つの貢献は，声部ラベルの推定結果を評価するため，[22] で提案されたピアノ採譜の評価尺度を拡張し，新たな評価尺度を提案することである．

2. 関連研究

この章では，音響から楽譜へのピアノ採譜，音価推定，声部分離に関する既存手法についてまとめる．

2.1 音響から楽譜へのピアノ採譜

これまでに，楽譜を出力するピアノ採譜手法がいくつか提案されていて，多段処理から構成される手法 [13,20] が高い精度を実現している．Cogliati ら [13] は，ピアノ演奏 MIDI に対してリズム量子化と声部分離をして，ピアノ楽譜を出力する採譜手法を提案した．この手法は，Temperley [4] が提案した確率モデルによって，MIDI から推定された拍節，声部，和声の構造を利用する．柴田ら [20] は，多段処理によって音響信号からピアノの楽譜を推定する最先端のピアノ採譜手法を提案した．この手法では，畳み込みニューラルネットワーク（CNN）を使い，ピアノ演奏の音響信号から，音高，発音時刻，消音時刻，ベロシティからなる演奏 MIDI を推定する．推定された秒単位の発音時刻は，隠れマルコフモデル（HMM）によって量子化される．音価と声部ラベルが別々に推定された後，MuseScore 3 を使いピアノ楽譜を出力する．本研究では，深層学習に基づいて音価と声部ラベルを同時に推定し，音高や発音時刻の推定精度に比べて低い，音価と声部の推定精度の改善を目指す．

音響信号から楽譜を直接推定するエンドツーエンドの採譜手法も提案されている．Carvalho ら [23] は，音響信号から Lilypond で表現されたピアノ楽譜を推定する seq2seq モデルを提案した．Román ら [24] は，**kern 形式で表現された楽譜を推定する畳み込み再帰型ニューラルネットワーク (CRNN) を提案した．このネットワークは，CTC 損失関数を使って学習される．これらのエンドツーエンドの手法は，短い音響信号あるいは合成された音響信号のみを使って評価されていて，実際の演奏に対する性能は評価されていない．

2.2 音価推定

音価と演奏された音の長さはいつも対応しているわけではないため，音価推定は難しい問題である [21]．Temperley [4] は，確率モデルに基づくリズム量子化手法を提案した．この手法はビートの位置を推定することで，発音時刻を量子化する．声部が推定された後，各音符の消音時刻（発音時刻と音価の和）は，同じ声部の次の音符の発音時刻に設定される．この手法の一つの問題点として，楽譜の読みやすさに影響する休符が，一つも出力されないことが挙げられる．中村ら [21] はマルコフ確率場に基づく音価推定手法を提案した．この手法は，音高と発音時刻が与えられたときの音価の確率分布を表現する文脈モデルと，音価から演奏された音の長さを生成する演奏モデルによって構成される．実験により，演奏モデルの効果が限定的であることが明らかになった [21]．したがって，本研究では音高と発音時刻の情報のみから音価を推定し，演奏された音の長さは利用しない．

2.3 声部分離

声部分離は，入力の音符列を音楽的な音の流れを表す複数のグループ（声部）に分離する問題である．Karydis ら [15] は，ピアノ楽譜に対するルールベースの声部分離手法を提案した．この手法は，発音時刻と音価が等しい音符をまとめる垂直方向の統合と，発音時間と音高が近い音符同士をまとめる水平方向の統合からなる．この手法はホモフォニックな声部に対応できるが，他の多くの声部分離の手法はモノフォニックな声部のみを推定する．McLeod ら [18] は，HMM を使い，MIDI に対する声部分離手法を提案し，高精度を実現した．Valk ら [19] は，DNN に基づく声部分離手法を提案した．この手法は深層順伝播型ニューラルネットワークを使い，33 個の特徴量によって表現された各音符を 5 つのクラスに分類する．ピアノ採譜では，これらの既存の声部分離手法は適切ではない．なぜなら，各声部は和音を含み，音価は正確に推定されないからである．また，楽譜を出力するには両手パートや声部ラベルが必要であり，クラスタリングではなく，クラス分類問題を解く必要がある．例えば，声部ラベルは各音符の旗の向きを決めるために必要となる．柴田ら [20] は，損失関数に基づく声部分離手法を提案した．この手法はピアノ採譜に適切な手法ではあるが，精度改善の余地があるため，本研究では深層学習に基づいて，ピアノ採譜に適切な声部分離手法を提案する．

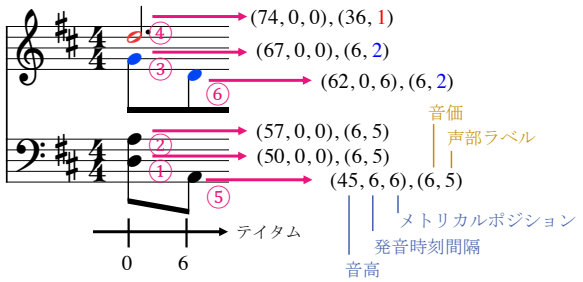

図 2 BiLSTM の入出力のデータ表現

3. 提案手法

この章では，音価と声部ラベルを同時に推定する BiLSTM ネットワーク，推定された音価を訂正する後処理，演奏不可能な音符を削除する後処理，データ拡張について述べる．

3.1 問題設定

ピアノ譜の各音符 $\mathbf{z}_n = (p_n, o_n, d_n, v_n)$ を音高 p_n，発音時刻 o_n，音価 d_n，声部ラベル v_n で表す．音高 $p_n \in \{0, \ldots, 127\}$ は MIDI ノートナンバーで表す（$60 = \mathrm{C4}$）．発音時刻 $o_n \geq 0$ と音価 $d_n \in \{0, \ldots, 479\}$ は，1 小節を 48 分割し，整数で表す．拍子が異なると時間分解能が異なることを注意する：例えば，四分音符は，4/4 拍子のとき 12 で表され，3/4 拍子のとき 16 で表される．MuseScore 3 や Finale といった楽譜編集ソフトウェアで採用されている慣習に従い，各手パートの最大声部数は 4 とする．声部ラベル $v_n = 1, 2, 3, 4$ は右手パートの声部を表し，$v_n = 5, 6, 7, 8$ は左手パートの声部を表す．ピアノ譜は発音時刻が早く，音高が低い音符から順に並べた音符列 $\mathbf{Z} = \{\mathbf{z}_n\}_{n=1}^N$ として表す．本研究の目標は，与えられた音高と発音時刻の列 $\{(p_n, o_n)\}_{n=1}^N$ から，音価と声部ラベルの列 $\{(d_n, v_n)\}_{n=1}^N$ を推定することである．

3.2 BiLSTM ネットワーク

音高と発音時刻から音価と声部ラベルを推定する BiLSTM ネットワークを提案する．まず，発音時刻 o_n を 1 つ前の音符の発音時刻との間隔 $i_n \in \{0, \ldots, 767\}$ と，メトリカルポジション $b_n \in \{0, \ldots, 47\}$ で表す：

$$i_n = o_n - o_{n-1}, \quad b_n = o_n \bmod 48. \tag{1}$$

このとき，ネットワークの入力は $\mathbf{X} = \{(p_n, i_n, b_n)\}_{n=1}^N$ であり，出力は $\mathbf{Y} = \{(d_n, v_n)\}_{n=1}^N$ である (図 2)．

提案するネットワークアーキテクチャを図 3(a) に示す．入力 $\mathbf{X}$ の各音符は $(128 \times 768 \times 48)$ 次元のワンホットベクトルで表現される．これらのワンホットベクトルは最初，全結合層によって 25 次元の特徴ベクトルに変換される．次に BiLSTM 層によって 50 次元のベクトル（潜在表現）に変換される．最後に，全結合層とソフトマックス層によって，音価の確率分布 $\boldsymbol{\pi}_n(\mathbf{X}) = \{\pi_n(d; \mathbf{X})\}_{d=0}^{479}$ と声部ラベルの確率分布 $\boldsymbol{\phi}_n(\mathbf{X}) = \{\phi_n(v; \mathbf{X})\}_{v=1}^{8}$ がそれぞれ出力される．ここで，$\pi_n(d; \mathbf{X})$ は n 番目の音符の音価が d である確率を表し，$\phi_n(v; \mathbf{X})$ は n 番目の音符の声部ラベルが v である確率を表す．

以下で定義されるクロスエントロピー損失関数を最小化することによって，ネットワークを学習する：

$$\mathcal{L} = \mathcal{L}_\mathrm{d} + \mathcal{L}_\mathrm{v}, \tag{2}$$

$$\mathcal{L}_\mathrm{d} = -\sum_{n=1}^{N} \log \pi_n(d_n^*; \mathbf{X}), \tag{3}$$

$$\mathcal{L}_\mathrm{v} = -\sum_{n=1}^{N} \log \phi_n(v_n^*; \mathbf{X}). \tag{4}$$

ここで，d_n^* と v_n^* はそれぞれ，正解の音価と声部ラベルを表す．音高と発音時刻 $\mathbf{X}$ が与えられたとき，音価と声部ラベルを次のように推定する：

$$\hat{d}_n = \arg\max_d \pi_n(d; \mathbf{X}), \tag{5}$$

$$\hat{v}_n = \arg\max_v \phi_n(v; \mathbf{X}). \tag{6}$$

ここで，$\hat{d}_n$ と $\hat{v}_n$ はそれぞれ，推定された音価と声部ラベルを表す．

3.3 ネットワークアーキテクチャの検討

図 3(a) に示されているネットワークアーキテクチャは，音価と声部ラベルの確率分布を均等に扱ったネットワークである．このネットワークを SIM (simultaneous) とする．本研究では，図 3 に示されている他のネットワークアーキテクチャを検討する．1 章で議論されたように，声部構造は音価を決定するのに重要な役割を果たす．この関係性を明示的に反映するため，2 つ目のネットワークアーキテクチャ (VLF; voice label first) を提案する．このネットワーク (図 3(b)) では，声部ラベルが初めに推定され，音価は声部ラベルを推定する際に使用した潜在表現を利用して推定される．比較のために，音価と声部の推定順序を反対にした 3 番目のネットワークアーキテクチャ (NVF; note value first) も検討する（図 3(c)）．以上 3 つのネットワーク，SIM, VLF, NVF は，式 (2) で定義される損失関数 $\mathcal{L}$ を最小化することで，マルチタスク学習の枠組みで学習される．マルチタスク学習の効果を検証するため，音価と声部を独立に推定する 4 つ目のネットワークアーキテクチャ (IND; independent) を検討する（図 3(d)）．IND は 2 つの BiLSTM ネットワークからなり，損失関数 $\mathcal{L}_\mathrm{d}$ と $\mathcal{L}_\mathrm{v}$ をそれぞれ最小化することで，別々に学習される．全てのネットワークアーキテクチャにおいて，最初の全結合層は 25 次元のベクトルを出力し，各 BiLSTM 層は各時刻で，50 次元の隠れベクトルを出力する．

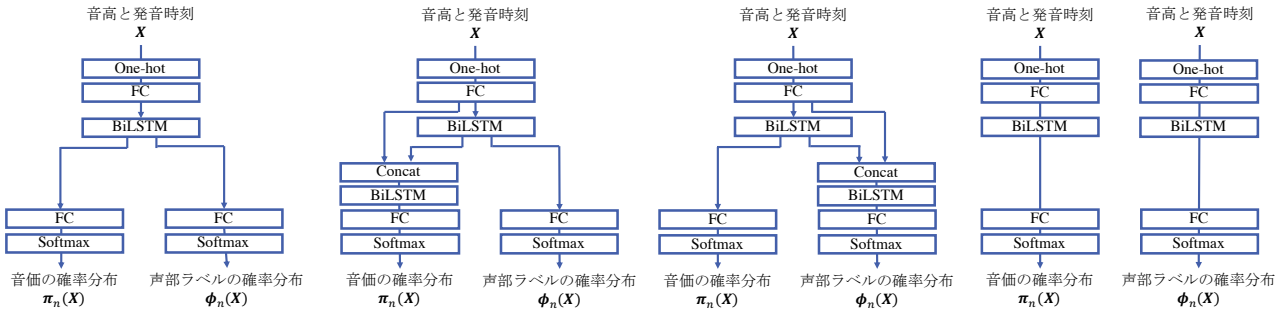

図 3 音価と声部ラベルを推定する BiLSTM ネットワーク．各ネットワークアーキテクチャは 3.3 節で説明される．

3.4 推定された音価に対する後処理

ネットワークによって推定された音価と声部ラベル $\{(\hat{d}_n, \hat{v}_n)\}_{n=1}^N$ は音楽的な慣習と矛盾することがある．一般的な制約として，発音時刻と声部が同じ音符は音価が等しいこと，消音時刻は同じ声部の次の発音時刻を超えないことが挙げられる．一つの声部に含まれる 2 つの音符 n と m について，これらの制約は次のように表現される：

$$o_n = o_m \implies d_n = d_m, \tag{7}$$

$$o_n < o_m \implies d_n \leq o_m - o_n. \tag{8}$$

これらの制約を満たすように，推定された音価 $\{\hat{d}_n\}_{n=1}^N$ を訂正するための 3 つの後処理手法を検討する．$\{n_k\}_{k=1}^K$ を発音時刻と声部が同じ音符とする．1 つ目の手法 (PP1) では，[13] と同じように，音価 $\{\hat{d}_{n_k}\}_{k=1}^K$ を次の発音時刻までの間隔に設定する．この方法では，音価は推定された声部ラベルのみから決まり，推定された音価の確率分布は使われていないことを注意する．2 つ目の手法 (PP2) では，音価を推定された音価の最大値に設定する：

$$\hat{d}'_{n_k} = \max_{l=1,\ldots,K} \hat{d}_{n_l}. \tag{9}$$

訂正された音価 $\hat{d}'_{n_k}$ が次の発音時刻までの間隔よりも大きい場合は，その間隔に修正する．3 つ目の手法 (PP3) では，制約を満たす音価の中から，確率の積が最大となるものを求める：

$$\hat{d}'_{n_k} = \arg\max_{d:\ d \leq d'} \prod_{l=1}^{K} \pi_{n_l}(d; \mathbf{X}). \tag{10}$$

ここで，d' は次の発音時刻までの間隔を表す．

3.5 演奏不可能な音符を削除する後処理

一般的なピアノの楽譜において，各手で同時に演奏される音符数は最大 5 つであり，音高の幅は 1 オクターブ以内である．しかし，あらかじめ推定された音高と発音時刻に誤りがある場合や，声部の推定結果に誤りがある場合に，採譜結果がこれらの演奏可能性の制約を満たさないことがある．そこで，各手パートの音符を必要最低限削除し，演奏可能な楽譜に訂正する後処理（DEL）を提案する．

DEL では，まず提案した BiLSTM ネットワークを使い，各音符の声部ラベルを推定し，入力の音符列を右手パートと左手パートに分離する．次に，各手パートの同時発音の音符のグループに対して，演奏可能になるまで音符を削除する．具体的には，音符のグループが演奏不可能の場合，音符数を 1 つ減らした全ての音符の組み合わせの中で，演奏可能なものがあればその中から，各音符の声部ラベルの確率の積が最大となる音符の組み合わせを選択する．この操作を演奏可能な音符のグループが見つかるまで続ける．

3.6 データ拡張

本研究では，多重音検出と発音時刻量子化の手法によって，予め推定された音高と発音時刻 $\mathbf{X}$ を入力とする．結果として，入力はテンポ推定の誤り，挿入・削除音符を含む．提案するネットワークをこれらの誤りに頑健にするため，学習データ $\mathcal{D}$ にテンポ誤り，挿入・削除音符を追加するデータ拡張を提案する．発音時刻の量子化において，半テンポ誤り・倍テンポ誤りが含まれることが多いため [20]，正解楽譜の発音時刻と音価を 2 倍あるいは半分にすることで，テンポを変換したデータセット $\mathcal{D}_{\mathrm{t}}$ を作る．また，多重音検出において，正解の音符からオクターブだけ音高がずれた音符が挿入されることが多くある．そこで，正解の音符をランダムに削除し，正解の音符から音高が 1 オクターブ異なる音符をランダムに挿入することで，挿入・削除音符を含むデータセット $\mathcal{D}_{\mathrm{em}}$ を作る．さらに，学習データの量を増やすため，他のデータ拡張手法を用いる．移調されたピアノ楽譜も音楽的に妥当な楽譜であるという仮定のもと，元の楽譜の音高を全て δ 半音だけスライドしたデータを使ってネットワークを学習する ($\delta = -12, -11, \ldots, 12$)．

4. 評価実験

この章では，提案手法の採譜精度を評価するための評価実験について報告する．

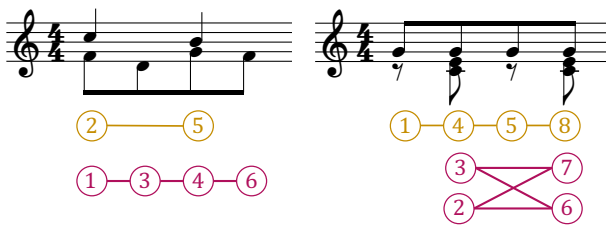

図 4 グラフによる声部構造の表現

4.1 実験設定

提案手法の実用的な状況下での精度を評価するため，既存の音響から楽譜への採譜システムと提案手法を統合し，評価用のピアノ演奏の音響信号を採譜した．まず，最先端の採譜手法 [20] で使われている多重音検出と発音時刻量子化の手法によって，音高と量子化された発音時刻を推定した．その推定結果である量子化 MIDI に対して，提案手法によって音価と声部ラベルを推定した．最後に楽譜編集ソフト MuseScore 3 を使って，MusicXML 形式の楽譜を出力した (図 1)．比較のために，既存の音価と声部の推定手法 [13, 20] による採譜も行った．手法 CTD16 [13] は，音価と声部ラベルの推定のために，Melisma Analyzer [4] を使用している．手法 SNY21 [20] は，現在最も精度の高い手法であり，統計モデルに基づく音価推定と動的計画法に基づく声部分離からなる．

評価データとして，MAPS-ENSTDkCl データセット [25] に含まれる，30 曲のクラシック曲のピアノ演奏データ，[20] でも使われた 81 曲のポピュラー曲のピアノカバー演奏データを使用した．これらの演奏データに対する正解楽譜を，MusicXML 形式で用意した．BiLSTM ネットワークの学習データとして，[20] でも使われた 80 曲のクラシック曲の楽譜と，763 曲のポピュラー音楽の楽譜を使用した．これらの学習データに対して，3.6 節で説明したデータ拡張手法を適用した．

4.2 評価尺度

採譜結果の評価尺度として，編集距離に基づく誤り率 [22] を採用した．この尺度は，音高誤り率 $\mathcal{E}_{\mathrm{p}}$，削除誤り率 $\mathcal{E}_{\mathrm{m}}$，挿入誤り率 $\mathcal{E}_{\mathrm{e}}$，発音時刻誤り率 $\mathcal{E}_{\mathrm{on}}$，消音時刻誤り率 $\mathcal{E}_{\mathrm{off}}$ から構成される．これらは，推定結果の楽譜と正解の楽譜のアラインメントをとってから計算される．このうち，消音時刻誤り率 $\mathcal{E}_{\mathrm{off}}$ を音価の推定精度を評価するために使う．この採譜尺度では，声部の推定精度を評価することができない．本研究では，推定された声部ラベルの誤り率 $\mathcal{E}_{\mathrm{v}}$ も計算する．平均誤り率 $\mathcal{E}_{\mathrm{all}}$ はこれら 6 つの誤り率の平均として計算される．

また，声部分離の推定精度を評価するため，声部分離の研究で従来使われている F 値 [17] を使用する．既存の評価尺度 [17] は，モノフォニックの声部を評価することを目的としている．本研究ではホモフォニックの声部も評価できるように拡張する．声部構造は，1 つの声部内の連続する和音の各音符を辺で繋いだグラフとして表現できる (図 4)．このグラフは隣接行列 (a_{ij}) によって表現できる：$a_{ij}=1$ は j 番目の音符は i 番目の音符と同じ声部に属し，i 番目の音符が構成する和音の次の和音に含まれることを表す．それ以外のとき，$a_{ij}=0$ である．(a_{ij}) を正解楽譜に対する隣接行列とし，$(\hat{a}_{ij})$ を推定された楽譜に対する隣接行列とする．適合率 $\mathcal{P}_{\mathrm{v}}$，再現率 $\mathcal{R}_{\mathrm{v}}$，F 値 $\mathcal{F}_{\mathrm{v}}$ を次のように定義する：

$$\mathcal{P}_{\mathrm{v}}=\frac{\sum_{i<j}a_{ij}\hat{a}_{ij}/\hat{w}_i}{\sum_{i<j}\hat{a}_{ij}/\hat{w}_i},\quad \mathcal{R}_{\mathrm{v}}=\frac{\sum_{i<j}a_{ij}\hat{a}_{ij}/w_i}{\sum_{i<j}a_{ij}/w_i},\quad (11)$$

$$\mathcal{F}_{\mathrm{v}}=\frac{2\mathcal{P}_{\mathrm{v}}\mathcal{R}_{\mathrm{v}}}{\mathcal{P}_{\mathrm{v}}+\mathcal{R}_{\mathrm{v}}}.\quad (12)$$

ここで，$\sum_{i<j}$ は全ての音符 i と，i の後ろに現れる全ての音符 j に関する和を意味する．各音符 i に対して，重み $w_i, \hat{w}_i$ を次のように定義する：

$$w_i=\sum_{j>i}a_{ij},\quad \hat{w}_i=\sum_{j>i}\hat{a}_{ij}.\quad (13)$$

この重みは声部の推定精度を，各和音に対して音符数にかかわらず均一に評価するために導入した．

表 1 MAPS データにおける音価と声部ラベルの誤り率と精度 (%)

手法	$\mathcal{E}_{\mathrm{off}}$	$\mathcal{E}_{\mathrm{v}}$	$\mathcal{P}_{\mathrm{v}}$	$\mathcal{R}_{\mathrm{v}}$	$\mathcal{F}_{\mathrm{v}}$
SIM+DA	33.3	**39.1**	63.9	**64.9**	64.0
VLF+DA	**32.2**	**39.0**	**65.2**	**65.7**	**65.1**
NVF+DA	**32.9**	40.7	63.1	62.6	62.5
IND+DA	**32.9**	40.5	64.1	63.8	63.6
VLF	**33.1**	**39.1**	**64.3**	64.3	64.0

表 2 J-pop データにおける音価と声部ラベルの誤り率と精度 (%)

手法	$\mathcal{E}_{\mathrm{off}}$	$\mathcal{E}_{\mathrm{v}}$	$\mathcal{P}_{\mathrm{v}}$	$\mathcal{R}_{\mathrm{v}}$	$\mathcal{F}_{\mathrm{v}}$
SIM+DA	**17.9**	12.2	**87.1**	**87.3**	**87.1**
VLF+DA	**17.2**	**11.4**	**87.7**	**87.7**	**87.6**
NVF+DA	**18.1**	12.4	**87.4**	**87.2**	**87.2**
IND+DA	18.7	12.5	**86.8**	86.4	86.5
VLF	**17.5**	**11.4**	**87.5**	**87.8**	**87.6**

表 3 異なる後処理に対する音価の誤り率 $\mathcal{E}_{\mathrm{off}}$ (%)

手法	MAPS	J-pop
VLF+DA	32.2	17.2
VLF+DA+PP1	**28.0**	**15.3**
VLF+DA+PP2	31.4	16.3
VLF+DA+PP3	32.2	16.8

4.3 実験結果

まず，4 つのネットワークアーキテクチャ (SIM, VLF, NVF, IND) の比較と，データ拡張の有効性の検証を行う．MAPS データセットと J-pop データセットに対する評価結果をそれぞれ，表 1 と表 2 に示す．データ拡張をした上で学習した 4 つのネットワークのうち，音価と声部ラベルの

表 4　採譜結果の誤り率と精度 (%)

手法	評価データ	$\mathcal{E}_{\mathrm{p}}$	$\mathcal{E}_{\mathrm{m}}$	$\mathcal{E}_{\mathrm{e}}$	$\mathcal{E}_{\mathrm{on}}$	$\mathcal{E}_{\mathrm{off}}$	$\mathcal{E}_{\mathrm{v}}$	$\mathcal{E}_{\mathrm{all}}$	$\mathcal{P}_{\mathrm{v}}$	$\mathcal{R}_{\mathrm{v}}$	$\mathcal{F}_{\mathrm{v}}$
提案 (VLF+DA+PP1)	MAPS	**0.67**	**8.11**	**6.23**	**11.6**	**28.0**	**39.1**	**15.6**	**65.2**	**65.7**	**65.1**
提案 (VLF+DA+PP1+DEL)	MAPS	**0.82**	9.37	**5.88**	**12.0**	**27.3**	**38.9**	**15.7**	**65.6**	**65.7**	**65.3**
SNY21 [20]	MAPS	**0.67**	**8.11**	**6.23**	**11.5**	28.3	44.6	16.6	62.4	59.4	60.6
CTD16 [13]	MAPS	**0.88**	13.5	**6.33**	16.8	44.0	74.3	26.0	56.0	42.5	47.9
提案 (VLF+DA+PP1)	J-pop	**0.61**	**4.03**	**7.29**	**2.67**	**15.3**	**11.4**	**6.89**	**87.6**	**87.7**	**87.6**
提案 (VLF+DA+PP1+DEL)	J-pop	**0.73**	**4.13**	**7.03**	**2.69**	**15.2**	**11.4**	**6.85**	**87.8**	**87.8**	**87.7**
SNY21 [20]	J-pop	**0.61**	**4.03**	**7.29**	**2.69**	20.9	18.0	8.92	78.6	77.0	77.7
CTD16 [13]	J-pop	**0.82**	12.8	**7.21**	8.48	55.7	65.8	25.1	51.3	38.8	44.0

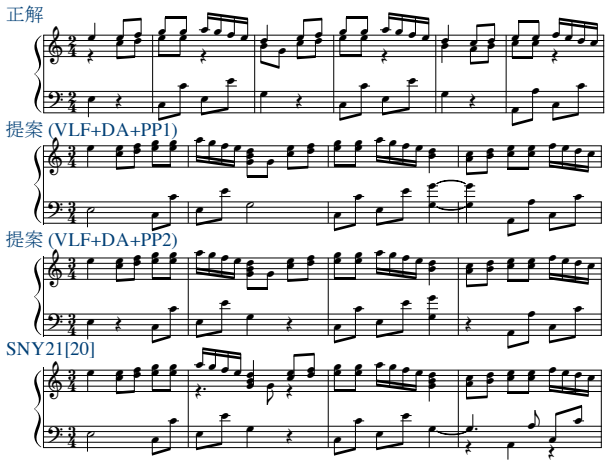

図 5　採譜結果の例

両方において，VLF が最も高い精度を実現した．NVF に比べて VLF の方が精度が高く，声部ラベルを先に推定することが有効であることがわかる．VLF と IND を比較することで，マルチタスク学習の効果が確認できる．VLF をデータ拡張した学習データと元々の学習データで学習した結果を比較することで，データ拡張の効果が確認できる．

次に，3 つの後処理手法を比較する（表 3）．1 番目の手法 (PP1) が最も低い誤り率を実現した．2 番目の手法 (PP2) は後処理前に比べてわずかに音価の精度を改善した．3 番目の手法 (PP3) はほとんど精度を改善していない．1 番目の手法では，音価は推定された声部ラベルのみから計算されていて，ネットワークが推定した音価の確率分布は使用していない．ここで，音価の推定が必要ないというわけではないことを注意する．音価推定と声部分離をマルチタスク学習することで，声部の推定精度が改善し，結果として音価の推定精度が改善しているからだ．

1 番目の後処理手法は休符を出力しないという制限がある．休符はアーティキュレーションを表現したり，楽譜の可読性を向上したりするために必要である．採譜結果の例を図 5 に示す．この例では，2 番目の後処理手法が正しく休符を推定できている．今後は，音価の推定精度を改善するために，休符の推定も重要になる．

最後に提案手法と既存の採譜手法を比較する [13, 20]．表 4 に，MAPS データと J-pop データそれぞれに対する採譜誤り率と声部分離の F 値を示す．どちらのデータセットに対しても，提案手法が最も高い精度を実現していることがわかる．MAPS データセットに対する精度の方が，J-pop データセットに対する精度よりも低いのは，クラシック曲の方がより複雑な声部構造を持っていること，学習データのクラシック曲が少ないことが原因であると考えられる．提案手法の後処理（DEL）によって，入力の音符のうち MAPS データセットに対して平均 3.5%，J-pop データセットに対して平均 1.4% の音符が削除された．この結果，削除誤り率 $\mathcal{E}_{\mathrm{m}}$ は増加し，挿入誤り率 $\mathcal{E}_{\mathrm{e}}$ は減少している．DEL の前後で平均誤り率 $\mathcal{E}_{\mathrm{all}}$ はほとんど変化がないが，採譜結果の演奏可能性が改善するため，DEL は有効である．

5. おわりに

本稿では深層学習に基づいて，事前に推定された音高と発音時刻から音価と声部を同時に推定する手法を提案した．音価と声部には相互関係があるため，BiLSTM ネットワークをマルチタスク学習の枠組みに従い学習した．実験により，提案手法は最新の多重音検出と発音時刻量子化の手法と統合することで，音響から楽譜へのピアノ採譜において，最先端の精度を実現できることを確認した．

音価と声部の推定精度は他の精度に比べて依然として低い．今後の課題として，データ表現やネットワークアーキテクチャをさらに検討し，推定された音価と声部の一貫性を改善する予定である．また，休符を正確に予測し，可読性の高い楽譜を出力するため，より洗練された後処理法や，演奏された音の長さの利用を検討する．

本研究では音価と声部の推定に焦点を当てたが，音価と声部の推定結果は，前段処理である発音時刻の量子化における誤りの影響を受けている．したがって，発音時刻の量子化を統合することが重要である．完全にエンドツーエンドでピアノ採譜を行うのは現状困難であるため [23, 24]，図 1 の多段処理を一つずつ統合していくことが有効である．

謝辞　本研究の一部は，JSPS 科研費 No. 19H04137, No. 20K21813, No. 21K12187, JST さきがけ No. JPMJPR20CB, および 2021 年度京都大学リサーチ・ディベロップメントプログラム【いしずえ】の支援を受けた．

参考文献

[1] 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).
[2] Desain, P. and Honing, H.: The quantization of musical time: A connectionist approach, *Computer Music Journal*, Vol. 13, No. 3, pp. 56–66 (1989).
[3] Raphael, C.: A Hybrid Graphical Model for Rhythmic Parsing, *Artificial Intelligence*, Vol. 137, pp. 217–238 (2002).
[4] Temperley, D.: A unified probabilistic model for polyphonic music analysis, *Journal of New Music Research*, Vol. 38, No. 1, pp. 3–18 (2009).
[5] Vincent, E., Bertin, N. and Badeau, R.: Adaptive harmonic spectral decomposition for multiple pitch estimation, *IEEE TASLP*, Vol. 18, No. 3, pp. 528–537 (2010).
[6] Benetos, E. and Weyde, T.: An efficient temporally-constrained probabilistic model for multiple-instrument music transcription, *ISMIR*, pp. 701–707 (2015).
[7] Sigtia, S., Benetos, E. and Dixon, S.: An end-to-end neural network for polyphonic piano music transcription, *IEEE/ACM TASLP*, Vol. 24, No. 5, pp. 927–939 (2016).
[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, *ISMIR*, pp. 50–57 (2018).
[9] Wu, Y.-T., Chen, B. and Su, L.: Polyphonic music transcription with semantic segmentation, *ICASSP*, pp. 166–170 (2019).
[10] Kong, Q., Li, B., Chen, J. and Wang, Y.: GiantMIDI-Piano: A large-scale MIDI dataset for classical piano music, *arXiv preprint arXiv:2010.07061* (2020).
[11] Nakamura, E., Yoshii, K. and Sagayama, S.: 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).
[12] Nakamura, E. and Yoshii, K.: Music Transcription Based on Bayesian Piece-Specific Score Models Capturing Repetitions, *arXiv preprint arXiv:1908.06969* (2019).
[13] Cogliati, A., Temperley, D. and Duan, Z.: Transcribing Human Piano Performances into Music Notation., *ISMIR*, pp. 758–764 (2016).
[14] Kilian, J. and Hoos, H. H.: Voice Separation-A Local Optimization Approach, *ISMIR*, pp. 39–46 (2002).
[15] Karydis, I., Nanopoulos, A., Papadopoulos, A., Cambouropoulos, E. and Manolopoulos, Y.: Horizontal and Vertical Integration/Segregation in Auditory Streaming: A Voice Separation Algorithm for Symbolic Musical Data, *Proc. of Sound and Music Computing Conference*, pp. 299–306 (2007).
[16] Cambouropoulos, E.: Voice and stream: Perceptual and computational modeling of voice separation, *Music Perception*, Vol. 26, No. 1, pp. 75–94 (2008).
[17] Duane, B. and Pardo, B.: Streaming from MIDI Using Constraint Satisfaction Optimization and Sequence Alignment., *Proc. of International Computer Music Conference* (2009).
[18] 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).
[19] de Valk, R. and Weyde, T.: Deep neural networks with voice entry estimation heuristics for voice separation in symbolic music representations, *ISMIR* (2018).
[20] Shibata, K., Nakamura, E. and Yoshii, K.: Non-Local Musical Statistics as Guides for Audio-to-Score Piano Transcription, *Information Sciences*, Vol. 566, pp. 262–280 (2021).
[21] Nakamura, E., Yoshii, K. and Dixon, S.: Note value recognition for piano transcription using Markov random fields, *IEEE/ACM TASLP*, Vol. 25, No. 9, pp. 1846–1858 (2017).
[22] Nakamura, E., Benetos, E., Yoshii, K. and Dixon, S.: Towards Complete Polyphonic Music Transcription: Integrating Multi-Pitch Detection and Rhythm Quantization, *ICASSP*, pp. 101–105 (2018).
[23] Carvalho, R. G. C. and Smaragdis, P.: Towards end-to-end polyphonic music transcription: Transforming music audio directly to a score, *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 151–155 (2017).
[24] Román, M. A., Pertusa, A. and Calvo-Zaragoza, J.: A holistic approach to polyphonic music transcription with neural networks, *ISMIR*, pp. 731–737 (2019).
[25] 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).