1P-01

ピアノ採譜のための深層学習に基づく音価と声部の同時推定

平松 祐紀　柴田 剛　錦見 亮　中村 栄太　吉井 和佳

京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

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．ピアノ採譜では音高とリズムの認識に加えて，楽譜表示のために各音符の両手パートと声部の推定が必要であり，これは楽譜の見やすさに影響する．ここで, 声部とは各手パートに含まれる複数の旋律を分けたパートである．例えば，右手パート内にメロディーとメロディーを支える副旋律が存在することがあり，これら２つは異なる声部に割り当てられる．

音高とリズムの認識は，フレーム単位で音高を推定する多重音検出と，リズムをテイタム単位に変換するリズム量子化という２つの問題に分けて研究されてきた．ピアノ採譜手法 [1] は，多重音検出，発音時刻の量子化と音価推定からなるリズム量子化，両手パート分離，声部分離といった多段処理で構成されている．楽譜の見やすさに影響する音価推定と声部分離はこれまで，マルコフ確率場に基づく音価推定手法 [2] や，ディープニューラルネットワーク（DNN）に基づく声部分離手法 [3] が提案されているが，両者は別々に扱われてきた．しかし，声部の定義は各音符の音価に依存し，多くの音符の音価は，同じ声部で異なる発音時刻を持つ次の音符との発音時刻の差になる，というように音価と声部には相互関係がある．

そこで，本研究では深層学習に基づいて，音価推定と声部分離を同時に行う手法を提案する．具体的には，音高と発音時刻の情報から音価と声部を同時に予測する双方向長短期記憶（Bi-LSTM）に基づくネットワークと，音価と声部の整合性に基づく音価の推定結果の訂正手法を提案する．

2. 提案法

本章では，Bi-LSTM に基づく音価と声部の同時推定手法と，音価と声部の整合性に基づく音価の推定結果の訂正手法について述べる（図 1）．

2.1 問題設定

推定されたピアノ譜の音高と発音時刻の列 $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ に対し，音価と声部の列 $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ を推定する．ここで，ピアノ譜の各音符は発音時刻が早く，音高が低いものから順に並べられ，N は音符数を表す．発音時刻と音価は 1 小節を 48 分割して表す．音高と発音時刻 $\mathbf{x}_n = (p_n, o_n, b_n)$ は，MIDI ノートナンバーで表される音高 $p_n \in \{0, \ldots, 127\}$，発音時刻と 1 つ前の音符の発音時刻との差 $o_n \in \{0, \ldots, 767\}$，発音時刻の小節内の相対位置 $b_n \in \{0, \ldots, 47\}$ の 3 つで表される．音価と声部 $\mathbf{y}_n = (d_n, v_n)$ は，継続時間 $d_n \in \{0, \ldots, 479\}$ と声部ラベル $v_n \in \{0, \ldots, 7\}$ の 2 つで表される．ここで，声部は各手パート 4 つを上限とし，声部ラベル v_n を 4 で割ったとき，商が両手パートを表し，余りが各パート内の声部を表す．

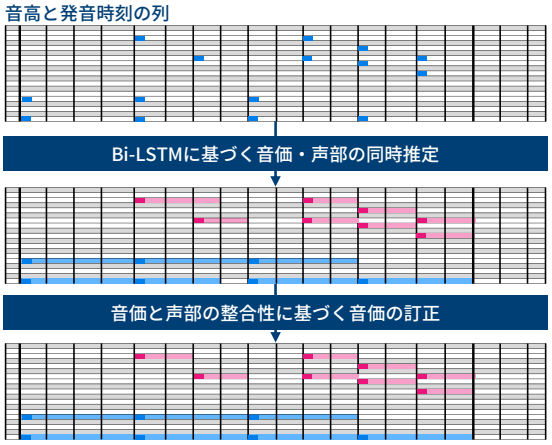

図 1: 音価と声部の同時推定

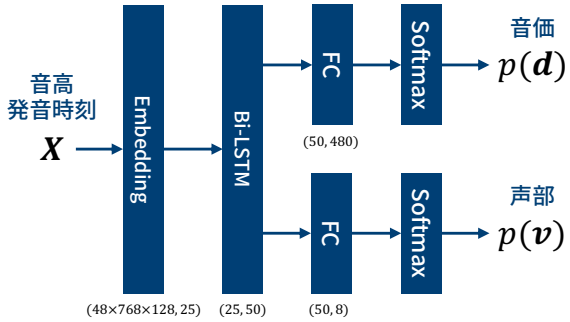

図 2: 音価と声部を同時推定するネットワーク

2.2 Bi-LSTM に基づく音価と声部の同時推定

音高と発音時刻の列 $\mathbf{X}$ に対して，音価と声部の列 $\mathbf{Y}$ を推定するネットワークを提案する．$\mathbf{X}$ が与えられたときの $\mathbf{Y}$ の条件付き確率分布 $p(\mathbf{Y}|\mathbf{X})$ を次のように分解する．

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{n=1}^{N} p(d_n|\mathbf{X}) p(v_n|\mathbf{X}) \tag{1}$$

提案するネットワークは入力 $\mathbf{X}$ に対し，音価 d_n の確率分布 $\{p(d_n = d|\mathbf{X})\}_{d=0}^{479}$ と声部 v_n の確率分布 $\{p(v_n = v|\mathbf{X})\}_{v=0}^{7}$ をそれぞれ出力する．入力の各音符 $\mathbf{x}_n$ を表すワンホットベクトルを，25 次元のベクトルに変換する埋め込み層，Bi-LSTM，音価と声部で異なる全結合層とソフトマックス層から構成される（図 2）．このネットワークは，以下で定義されるクロスエントロピーに基づく損失関数 $\mathcal{L}_d$ と $\mathcal{L}_v$ の和を最小化することにより学習する．

Joint Estimation of Note Values and Voice Parts Based on Deep Learning for Piano Transcription: *Yuki Hiramatsu, Go Shibata, Ryo Nishikimi, Eita Nakamura, and Kazuyoshi Yoshii (Kyoto Univ.)*

表 1: 実験結果（%）

手法	E_{p}	E_{m}	E_{e}	E_{on}	E_{off}	E_{all}	$\mathcal{F}_{\mathrm{p}}$	$\mathcal{F}_{\mathrm{voi}}$	$\mathcal{F}_{\mathrm{met}}$	$\mathcal{F}_{\mathrm{val}}$	$\mathcal{F}_{\mathrm{harm}}$	$\mathcal{F}_{\mathrm{MV2H}}$
Bi-LSTM	0.60	4.11	7.37	2.70	17.8	6.51	92.9	89.0	83.9	94.5	91.7	90.4
Bi-LSTM + 訂正	0.60	4.11	7.37	2.45	**16.0**	**6.11**	93.2	**89.7**	84.2	95.6	91.7	**90.9**
Shibata ら [1]	0.60	4.11	7.37	2.46	20.7	7.06	93.2	79.7	84.3	95.6	91.7	88.9

$$\mathcal{L}_d = -\sum_{n=1}^{N}\sum_{d=0}^{479} p(d_n = d|\mathbf{X}) \log p(d_n = d|\mathbf{X}) \quad (2)$$

$$\mathcal{L}_v = -\sum_{n=1}^{N}\sum_{v=0}^{7} p(v_n = v|\mathbf{X}) \log p(v_n = v|\mathbf{X}) \quad (3)$$

音価 d_n と声部 v_n の推定は，確率の最大化によって行う．

$$\hat{d_n} = \arg\max_{d} p(d_n = d|\mathbf{X}) \quad (4)$$

$$\hat{v_n} = \arg\max_{v} p(v_n = v|\mathbf{X}) \quad (5)$$

2.3 音価と声部の整合性に基づく音価の訂正

一つの声部に含まれる音符は，(1) 発音時刻が等しい音符の音価は等しく，(2) 消音時刻は，発音時刻と次の異なる発音時刻との間に存在する，という 2 つの条件を満たす必要がある．Bi-LSTM に基づく音価と声部の推定結果として，これらの制約を満たさないものが考えられる．そこで，推定された声部ごとに，発音時刻が同じ音符の音価を適切に揃えることで，これらの制約を満たすようにする．具体的には，推定された音価の最大値と，発音時刻と次の異なる発音時刻との差のうち，小さい方を音価の訂正結果とする．

3. 評価実験

実験には，文献 [1] で用いられた J-POP のピアノカバー 81 曲を使用した．ピアノ採譜手法 [1] で提案された，畳み込みニューラルネットワーク（CNN）に基づく多重音検出と，隠れマルコフモデル（HMM）に基づく発音時刻の量子化を行い，音高と発音時刻の列 $\mathbf{X}$ を得た．ネットワークの学習には，文献 [1] で使用されたピアノカバー 777 曲の楽譜を用いた．学習データの音高と発音時刻は，ピアノ採譜手法 [1] によって推定されたものではないことを注意する．ネットワークの汎化性能を高くするため，音高を最大 ±12 シフトすることで，学習データを 25 倍にした．

提案手法とピアノ採譜手法 [1] で使われた音価と声部の推定手法を比較した．採譜尺度として，[4] で提案された編集距離に基づく採譜尺度と，[5] で提案された MV2H を用いた．前者は，音高誤り率 E_{p}，削除誤り率 E_{m}，挿入誤り率 E_{e}，発音時刻の誤り率 E_{on}，消音時刻の誤り率 E_{off} と，これら 5 つの平均 E_{all} からなる．後者は，多重音検出の精度 $\mathcal{F}_{\mathrm{p}}$，声部分離の精度 $\mathcal{F}_{\mathrm{voi}}$，拍節配位の精度 $\mathcal{F}_{\mathrm{met}}$，音価推定の精度 $\mathcal{F}_{\mathrm{val}}$，和声解析の精度 $\mathcal{F}_{\mathrm{harm}}$ と，これら 5 つの平均 $\mathcal{F}_{\mathrm{MV2H}}$ からなる．

表 1 に，提案手法と採譜手法 [1] の評価値を示す．提案手法が消音時刻の誤り率 E_{off} と声部分離の精度 $\mathcal{F}_{\mathrm{voi}}$ を改善し，平均誤り率 E_{all} と平均精度 $\mathcal{F}_{\mathrm{MV2H}}$ も改善していることがわかる．図 3 に，音価と声部の推定結果の例を示す．赤く囲われた部分で，提案手法がメロディーの音価と声部を正しく推定していることがわかる．以上の結果から，提案手法によって音価と声部の推定精度が向上し，楽譜の見やすさを改善できることが確認できた．しかし，Bi-LSTM による推定結果から，音価と声部の整合性に基づいて音価が訂正された音符は，全体の 6.6% を占め，Bi-LSTM は音価と声部の整合性を十分に学習できていないことも確認できた．

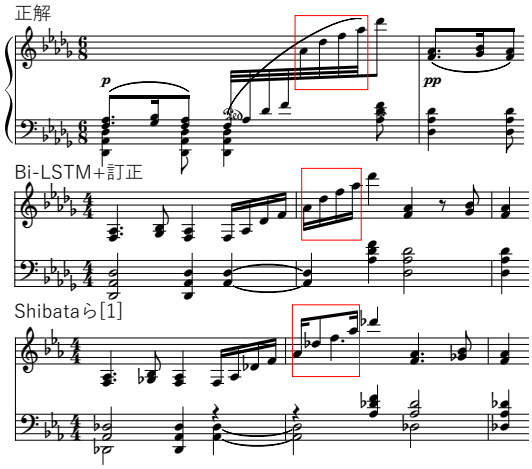

図 3: 推定結果の例

4. おわりに

本稿では，ピアノ採譜のための深層学習に基づく音価と声部の同時推定手法を提案した．実験の結果，提案手法によって音価と声部の推定精度が改善したが，ネットワークは音価と声部の整合性を十分に学習できていないことが確認できた．今後は，入力の音符列が誤りを含むことを考慮したネットワークの学習を行う予定である．また，入出力のデータ表現の検討や，音価と声部の整合性を評価する確率モデルの定式化と効率的に最適な系列を求めるアルゴリズムの導出も行う予定である．

謝辞 本研究の一部は，JST ACCEL No. JPM-JAC1602，JSPS 科研費 No. 16H01744, No. 19H04137, No. 19K20340 の支援を受けた．

参考文献

[1] K. Shibata *et al.*: “Non-Local Musical Statistics as Guides for Audio-to-Score Piano Transcription,” *arXiv:2008.12710*, 2020.

[2] E. Nakamura *et al.*: “Note Value Recognition for Piano Transcription Using Markov Random Fields,” *TASLP*, vol.25, no.9, 2017.

[3] R. de Valk *et al.*: “Deep Neural Networks with Voice Entry Estimation Heuristics for Voice Separation in Symbolic Music Representations,” *ISMIR*, 281–288, 2018.

[4] E. Nakamura *et al.*: “Towards Complete Polyphonic Music Transcription: Integrating Multi-Pitch Detection and Rhythm Quantization,” *ICASSP*, 101–105, 2018.

[5] A. McLeod *et al.*: “Evaluating Automatic Polyphonic Music Transcription,” *ICASSP*, 42–49, 2018.