Vol.2021-MUS-132 No.11
2021/9/16

難易度に応じた音符統計量と深層学習に基づくバンド譜のピアノ編曲

寺尾 萌夢[1,a)]　吉井 和佳[1,2,b)]

概要：本稿では，深層ニューラルネットワーク（DNN）を用いて，ポピュラー音楽のバンド譜を，指定された難易度（初級または上級）のピアノ譜に編曲する手法について述べる．バンド譜のピアノ編曲においては，ピアノ譜の持つ複雑な同時的・経時的構造を捉えた上で，音楽的に妥当なピアノ編曲は一意ではないことを考慮することが肝要である．本研究ではまず，バンド譜のメロディ（歌声）・伴奏パートおよびピアノ譜の右手・左手パートの関係を統計的に分析することにより，バンド譜に含まれる音符を上下 1 オクターブシフトして得られた拡張バンド譜から，音符を選択することでピアノ譜が得るという方針が妥当であることを示す．提案するピアノ編曲法の特徴は，拡張バンド譜の各音符を取捨選択する DNN を教師あり学習する際に，演奏の難易度に影響を与える大局的な音符統計量として，同時発音数・同時発音における音高幅・音符密度の分布に基づく正則化を導入することにある．これにより，DNN が出力するピアノ譜の音符配置を，ひとつの編曲例に過ぎない正解ピアノ譜と局所的に一致させるだけではなく，音符統計量の観点で好ましくなるよう大局的に最適化する学習を行う．実験により，大局的な音符統計量を考慮することの有効性を検証する．

1. はじめに

音楽情報処理分野では，楽器や音楽スタイルの変換，メロディからの伴奏生成など，様々な編曲タスクが研究されている．演奏楽器の変換では，特定の楽器で演奏されている楽曲を異なる楽器での演奏形態に変換する際に，曲のジャンルや対象となる楽器の特性などの専門知識が必要となる．そのため，この過程を自動化することは依然として難しい課題である．これまで，ピアノ編曲 [1–6] や, ギター編曲 [7–9], オーケストラへの編曲 [10,11] が行われている．本稿では，ポピュラー音楽の自動ピアノ編曲に取り組む．

自動ピアノ編曲には，隠れマルコフモデル (HMM) を用いて，入力された楽譜から音符を選択する手法 [5] が提案されている．この手法は統計的な枠組みに立脚しており，妥当なピアノ楽譜を得ることができるが，コード進行や声部など，演奏難易度に影響を与えるピアノ楽譜の複雑な同時的・経時的構造を十分に表現できるわけではなかった．

最近，深層ニューラルネットワーク (DNN) を用いて音楽（MIDI データ）の生成や編曲，スタイル変換を行う研究が盛んになりつつあるが [6]，統計的に妥当なピアノ譜を難易度別に推定しようとする試みは未だ行われていない．

ピアノ編曲の基本的な問題は，与えられた楽曲の「正解」となるピアノ譜が一意に定まらないことである．しかし，実際には無数にあるピアノ譜のうちの 1 つだけが正解として与えられる．一般的に，推定されたピアノ譜の品質は，推定ピアノ譜と正解ピアノ譜の音符単位での比較によって評価されるが，このような評価指標では，自動ピアノ編曲の可能性を制限してしまう．これは，DNN を音符単位で教師あり学習する際にも問題となる．

本研究ではまず，ポピュラー音楽のバンド譜と対応する難易度ラベル付きのピアノ譜を収集し，バンド譜のメロディ・伴奏パートとピアノ譜の右手・左手パートの関係（図 5）や，難易度毎のピアノ譜の特徴（図 4）を統計的に調査する．この結果に基づき，バンド譜に含まれる音符およびそれらをオクターブシフトした音符の集合から，音符を削除することでピアノ譜が得られるという妥当な仮定をおく．バンド譜のメロディ・伴奏パートからマスクを推定する DNN を難易度条件付きで学習する際には，推定されるピアノ譜が統計的に適切な性質を持つように誘導を行う．具体的には，正解のピアノ譜との誤差に加えて，実際のピアノ譜が持つ統計量との誤差を同時に最小化する．実験結果は，ピアノ譜の一貫性を向上させるために，大局的な音符統計量を考慮することの重要性を示している．

1 京都大学 大学院情報学研究科
2 科学技術振興機構 戦略的創造研究推進事業（さきがけ）
a) terao@sap.ist.i.kyoto-u.ac.jp
b) yoshii@i.kyoto-u.ac.jp

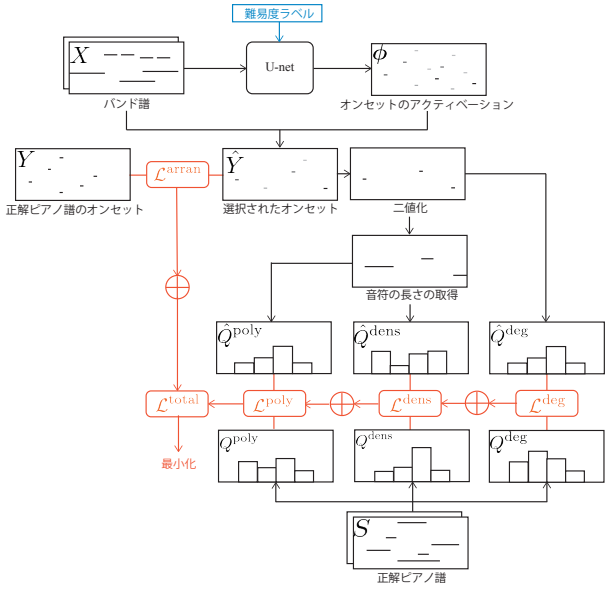

図 1 提案するピアノ編曲手法

2. 関連研究

ここでは, ピアノ編曲の手法を演奏難易度の定義, 音楽的要素の活用，スコアリダクションの観点から総括する．

2.1 演奏難易度の定義

ピアノ楽譜の難易度に関連する重要な概念として，ピアノ楽譜の演奏可能性が挙げられる．Chiuet ら [1] は，演奏可能性の指標として，1）同時に演奏されるフレーズは最大でも 5 つまでであること，2）すべての音符がピアノの音域内にあること，3）同時に演奏される音の最高音と最低音の距離が両手の指の届く範囲にあることを挙げている．Huang ら [3] は，ピアノ以外の楽器にも適用可能な同様の指標を提案しており，音程，同時発音数，楽器自体の音高の制約，音符の重なりに関する制約を考慮している．大沼ら [2] は，編曲家のピアノ編曲過程における問題とその解決の作業を分析し，指が届かない場合や，片手のパートの同時発音数が 6 個以上である場合に楽譜に手を加える必要があると報告している．中村ら [5] は，楽譜の演奏可能性は演奏者の技量や楽曲のテンポにも依存することを指摘している．そこで，演奏の難易度を定量化し，その難易度の違いを演奏間違いの予測能力で評価する手法を提案している．

2.2 音楽的要素の活用

Wang ら [6] は，ビート，コード，キーの情報や，複数のピアノ編曲案を含むポピュラー音楽データセット「Pop909」を提案し，トランスフォーマ・アーキテクチャ [12] を用いた多声部な音楽の生成と，メロディからの伴奏の生成について検討している．ピアノ編曲に関連した情報を与えることで，元のメロディと伴奏のハーモニーを生成された音楽に取り入れつつ，一貫したリズムパターンを生成することができる．高森ら [4] は，原曲の楽譜から抽出できるメロディ，コード，リズム，音符数を考慮したピアノ編曲法を提案している．右手パートではコードを構成する音から選択した音符をメロディに加え，左手パートではポピュラー音楽のピアノ譜からなる伴奏データベースから選択している．

2.3 スコアリダクション

Chiuet ら [1] は，重要な音楽フレーズをその有用性とピアノの演奏可能性に応じて選択するピアノ編曲法を提案している．中村ら [5] は，ピアノ楽譜の確率的生成モデルに基づくピアノ編曲手法 [13] を拡張し，HMM によって推定される運指の動き [14, 15] を取り入れている．

3. 提案手法

本章では，バンドのポピュラー音楽の楽譜を，指定した難易度に対応する音符の大域的な統計に着目して，ピアノスコアに編曲する手法を提案する (図 1).

3.1 問題設定

$\mathbf{B} \in \{0,1\}^{C_B \times P \times N}$ を，バンド譜，$\mathbf{S} \in \{0,1\}^{C_S \times P \times N}$ を対応するピアノ譜とする．$P = 128$ は MIDI フォーマットのピッチナンバー，N は 1 6 分音符単位での曲の長さを表す（テイタム数）．$C_B = 2$ と $C_S = 2$ はそれぞれ，$\mathbf{B}$ と $\mathbf{S}$ のチャンネル数であり，$\mathbf{B}$ の 2 つのチャンネルは，バンド譜のメロディ（歌声）パートと，伴奏パートを表している．同様に $\mathbf{S}$ の 2 つのチャンネルは，ピアノ譜の右手パート，左手パートを表している．$\mathbf{B}$ と $\mathbf{S}$ はピアノロールのフォーマットでスコアを表しており，ピッチ p の音符が n 番目のテイタムに存在するときに，$B_{c,p,n} = 1$ であり，存在しないときは $B_{c,p,n} = 0$ である．また，難易度ラベルとして，$L = \{0,1\}$ を導入する．$L = 0$ は初級，$L = 1$ は上級を表す．

$\mathbf{B}$ と L から，$\mathbf{S}$ を直接求める代わりに，マスク行列 $\boldsymbol{\phi} \in [0,1]^{C_\phi \times P \times N}$ を求め，さらにそこから推定ピアノ譜として $\hat{\mathbf{Y}} \in \{0,1\}^{C_{\hat{Y}} \times P \times N}$ を計算する．ここで，$C_\phi = 2$, $C_{\hat{Y}} = 2$ であり，2 つのチャンネルは順に，右手パートのオンセット，左手パートのオンセットを表している．$\mathbf{S}$ を得るには，$\mathbf{B}$ と L より，マスク行列 $\boldsymbol{\phi}$ を求めたあと，最終的に $\hat{\mathbf{Y}}$ から $\mathbf{S}$ を求める．

いま，バンド譜 $\mathbf{B}$ と難易度ラベル L が与えられたときに，$\mathbf{S}$ を推定するモデルを推定したい．まず，与えられたバンド譜からマスクを推定する．そのマスクを使い，バンド譜から音符を選択することで，ピアノ譜へ変換する．このモデルはバンド譜から選んできた音符と，ピアノ譜に関する統計量が正解ピアノ譜（人によって編曲されたピアノ譜）に近づくように訓練する．具体的には，音符単位での

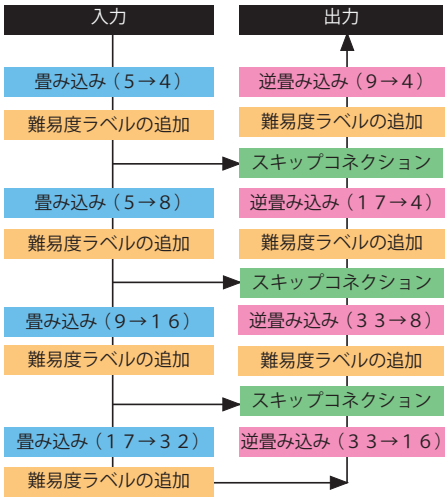

図 2 U-Net のアーキテクチャの詳細． かっこの中身はチャンネル数の変化を示している．

誤りに関する損失関数と，同時発音数，同時発音における音高幅，音符密度に関する損失関数を最小化する．

3.2 マスク推定

本研究では，バンド譜と難易度ラベルからマスク推定するモデルとして U-net [16] を使った (図 2)．U-Net は，$\mathbf{X} \in \{0,1\}^{C_X \times P \times N}$ を入力とし，$\boldsymbol{\phi}$ を出力する．$\mathbf{X}$ はチャンネル数が 5 $(C_X = 5)$ の行列で，バンド譜と難易度ラベルの情報を含んでいる．特に，$\mathbf{X}$ は，$\mathbf{B}$ と，$\mathbf{B}$ のオンセットを表す 2 値の行列, 難易度ラベル L を並べた行列 $\mathbf{L}$ から構成される．$\mathbf{X}$ の 5 つのチャンネルは順に，メロディパート，伴奏パート，メロディパートのオンセット，伴奏パートのオンセット，難易度ラベルを表す．U-Net の最終層にはシグモイド関数を使い，$\boldsymbol{\phi}$ は $[0,1]$ である．出力のマスクからピアノ譜 $\mathbf{S}$ を計算する．

バンド譜とそれに対応するピアノ譜のペアが与えられたとき，推定ピアノ譜の正しさを推定ピアノ譜と正解ピアノ譜の近さと，推定ピアノ譜と正解ピアノ譜の音符統計量の近さという，2 つの側面から評価する損失関数を最小化するように U-net が最適化される．このように損失関数は，音符単位のロスと，音符統計量のロスから構成される．音符単位のロスを最小化することで，推定ピアノ譜を正解ピアノ譜に近づけることが出来る．音符統計量のロスを最小化することで，推定ピアノ譜が一貫したものであり，難易度ラベルを正しく反映することが出来る．

3.3 音符単位のロス

バンド譜から音符を選択するため，バンド譜に含まれる音符及びそれらをオクターブシフトした音符の集合のオンセットを表した行列を $\mathbf{Z} \in \{0,1\}^{2 \times P \times N}$ とする．

$$
\begin{aligned}
Z_{0,p,n} =& Z_{1,p,n} \qquad (1)\\
=& \max\{X_{c,p,n}, \epsilon \times X_{c,p+j,n}, \mid j \in \{-12,12\}, c \in \{0,1\}\}
\end{aligned}
$$

ここで，$p \pm j < 1$ や，$P < p \pm j$ の場合は，$Z = 0$ である．また，ϵ は，上下 1 オクターブシフトされた音符の重要度のパラメータである．$Z_{2,p,n}$ と $Z_{3,p,n}$ は，必ずしもメロディパートと右手パート，伴奏パートと左手パートが対応しているわけではないので，メロディパートと伴奏パートのオンセットを両方含んでいる．これにより例えば，伴奏パートに含まれる音符を右手で演奏する，といったことが可能になる．

また，$C_Y = 2$ として，正解ピアノ譜のオンセットを表す行列を $\mathbf{Y} \in \{0,1\}^{C_Y \times P \times N}$ とする．$\boldsymbol{\phi}$ と $\mathbf{Z}$ の要素積が音符の選択に対応していることに注意して，以下の損失関数を最小化するように U-net を学習する．

$$
\begin{aligned}
\mathcal{L}^{\text{arran}}(\boldsymbol{\phi}) = -\sum_{c=1}^{2}\sum_{p=1}^{P}\sum_{n=1}^{N}\Big(& aY_{c,p,n}\log\phi_{c,p,n}Z_{c,p,n}\\
&+(1-Y_{c,p,n})\log\left(1-\log\phi_{c,p,n}Z_{c,p,n}\right)\Big) \qquad (2)
\end{aligned}
$$

ここで，$\alpha \in \mathbb{R}_+$ は正例に対する重みである．

3.4 音符統計量のロス

音符統計量のロスは推定ピアノ譜と正解ピアノ譜から得られた統計量の分布の距離として定義される．特に本研究では，同時発音数，同時発音における音高幅，音符密度の 3 つの統計量について考える．計算手順を以下に表す．（1）各テイタムごとに同時発音数，音高幅，各小節ごとに音符密度を計算する．（それぞれ，$\mathbf{C}^{\text{poly}}$，$\mathbf{C}^{\text{deg}}$，$\mathbf{C}^{\text{dens}}$ とする．）（2）$\mathbf{C}^{\text{poly}}$，$\mathbf{C}^{\text{deg}}$，$\mathbf{C}^{\text{dens}}$ の要素をカウントし，分布を計算する．

$\boldsymbol{\phi}^{\text{hard}}$ を，$\boldsymbol{\phi}$ を次の式で，閾値 0.5 で二値化したものとする．

$$
\phi^{\text{hard}}_{c,p,n} = \begin{cases} 1 & (\phi_{c,p,n} \geq 0.5) \\ 0 & (\phi_{c,p,n} < 0.5) \end{cases} \qquad (3)
$$

ピアノ譜の推定のため，$\boldsymbol{\phi} \odot \mathbf{Z}$ を微分可能な方法で二値化する必要があるので，次の Gumbel-sigmoid trick を使い，$\hat{\mathbf{Y}} \in \{0,1\}^{2 \times P \times N}$ を得る．

$$
\hat{\mathbf{Y}} = \boldsymbol{\phi}^{\text{hard}} \odot \mathbf{Z} - \boldsymbol{\phi}^{\text{detach}} \odot \mathbf{Z} + \boldsymbol{\phi} \odot \mathbf{Z} \qquad (4)
$$

ここで，$\boldsymbol{\phi}^{\text{detach}}$ は $\boldsymbol{\phi}$ の勾配を取り除いたものとする．$\odot$ はアダマール積を表している．

まず，各テイタムごとの情報が必要な $\mathbf{C}^{\text{poly}}$，$\mathbf{C}^{\text{deg}}$ を計算するため，オンセットのみを表す $\hat{\mathbf{Y}}$ とバンド譜から，選択された音符の長さを計算する必要がある．

いま，$\mathbf{A} \in \mathbb{N}^A$ をバンド譜 $\mathbf{B}$ に含まれる音符のピッチナンバーを並べた系列とする．A はバンド譜に含まれる音符

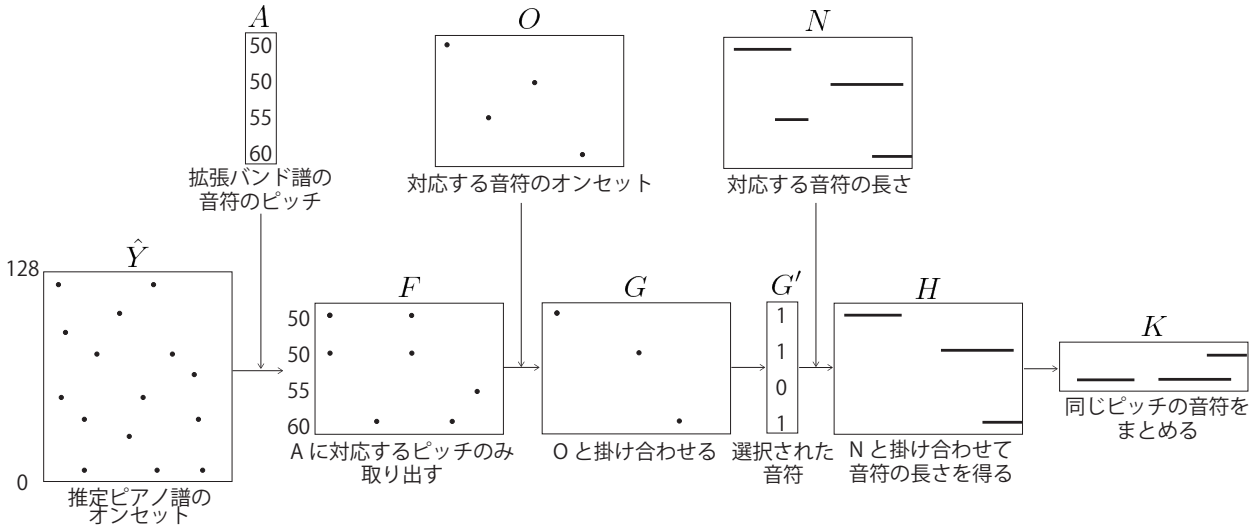

図 3 K の計算方法

の総数である．また，$\mathbf{N} \in \{0,1\}^{A\times N}$ を $\mathbf{A}$ の各音符の位置を表す行列とし，a 番目の音符が n 番目のテイタムに存在する場合，$N_{a,n}=1$ である．$\mathbf{O} \in \{0,1\}^{A\times N}$ は，$\mathbf{A}$ の各音符のオンセットを表す行列とし，a 番目の音符のオンセットが n 番目のテイタムに存在する場合，$O_{a,n}=1$ である．次に，補助的な行列として $\mathbf{F} \in [0,1]^{2\times A\times N}$ を定義する．

$$F_{c,a,n} = \{\hat{Y}_{c,A_a,n}\} \tag{5}$$

ここで，$\mathbf{F}$ の a 番目の行は $\hat{\mathbf{Y}}$ の $\mathbf{A}$ の a 番目のピッチの行と等しい．

$\mathbf{G} \in \{0,1\}^{C_{\hat{\mathbf{Y}}}\times A\times N}$ を $c \in 1,2$ として $\mathbf{F}_c$ と $\mathbf{O}$ のアダマール積を計算することで，各音符に対応する位置のオンセットのみを取り出すことが出来る．さらに $\mathbf{G}$ をテイタムの方向に足し合わせたものを，$\mathbf{G}'$ とすると，$\mathbf{G}'$ が 1 ならその音符を選択，0 なら削除したことになる．また，$\mathbf{H} \in \{0,1\}^{C_{\hat{\mathbf{Y}}}\times A\times N}$ を $c \in 1,2$ として $\mathbf{D}_n$ に $\mathbf{G}'_c$ を掛け合わせることで，選択された音符のみ長さを得る．

$$\mathbf{G}_c = \mathbf{F}_c \odot \mathbf{O} \tag{6}$$

$$\mathbf{G}' = \sum_{n=1}^{N} \mathbf{G}_n \tag{7}$$

$$\mathbf{H}_{c,n} = \mathbf{N}_n \odot \mathbf{G}'_c \tag{8}$$

$\mathbf{C}^{\mathrm{poly}}$ を計算するため，$\mathbf{H}$ をピッチが等しい音符に関してまとめた行列を $\mathbf{K} \in \{0,1\}^{C_{\hat{Y}}\times P\times N}$ とする (図 3)．

まず，推定ピアノ譜から $\hat{\mathbf{C}}^{\mathrm{poly}}$，$\hat{\mathbf{C}}^{\mathrm{deg}}$，$\hat{\mathbf{C}}^{\mathrm{dens}}$ を計算する．

$$\hat{C}^{\mathrm{poly}}_{c,n} = \sum_{p=1}^{P} K_{c,p,n} \tag{9}$$

$$\hat{C}^{\mathrm{deg}}_{c,n} = \max(\mathbf{H}_{c,n} \odot \mathbf{A}) - \min_{a\in J}(\mathbf{H}_{c,n} \odot \mathbf{A})_a \quad (J = \{a \mid (\mathbf{H}_{c,n} \odot \mathbf{A})_a > 0) \tag{10}$$

$$\hat{C}^{\mathrm{dens}}_{c,m} = \sum_{p=1}^{P}\sum_{n=1}^{16} \hat{Y}_{c,p,16m+n} \tag{11}$$

次に，正解ピアノ譜から $\mathbf{C}^{\mathrm{poly}}$，$\mathbf{C}^{\mathrm{deg}}$，$\mathbf{C}^{\mathrm{dens}}$ を計算する．

$$C^{\mathrm{poly}}_{c,n} = \sum_{p=1}^{P} S_{c,p,n} \tag{12}$$

$$C^{\mathrm{deg}}_{c,n} = \max(\mathbf{S}_{c,n} \odot \mathbf{p}) - \min_{p\in J}(\mathbf{S}_{c,n} \odot \mathbf{p})_p \quad (J = \{p \mid (\mathbf{S}_{c,n} \odot \mathbf{p})_p > 0) \tag{13}$$

$$C^{\mathrm{dens}}_{c,m} = \sum_{p=1}^{P}\sum_{n=1}^{16} Y_{c,p,16m+n} \tag{14}$$

ここで，m は小節数を表しており，n が 16 分音符数を表すことに注意すると，$1 \le m \le N/16$ である．$\mathbf{p} \in \mathbb{R}^P$ は要素にピッチ番号 $1 \le p \le P$ をもつベクトルとする．

次にこれらの C_n，$\hat{C}_n$ を用いて分布を計算する．以降では C_n を用いるが，$\hat{C}_n$ でも同様である．まず，$i \ge 0$ をヒストグラムの階級をとして，$i=k$ の時，($k-1$ の数) $\times 1+$ ($k-2$ の数) $\times 2 + \cdots +$ (0 の数) $\times k$ を計算する．

$$f_i(\mathbf{Y}) = \sum_{n=1}^{N} \mathrm{ReLU}\left(-C^{\mathrm{poly}}_n + i\right) \tag{15}$$

$$g_i(\mathbf{Y}) = \sum_{n=1}^{N} \mathrm{ReLU}\left(-C^{\mathrm{deg}}_n + i\right) \tag{16}$$

$$h_i(\mathbf{Y}) = \sum_{m=1}^{N/16} \mathrm{ReLU}\left(-C^{\mathrm{dens}}_m + i\right) \tag{17}$$

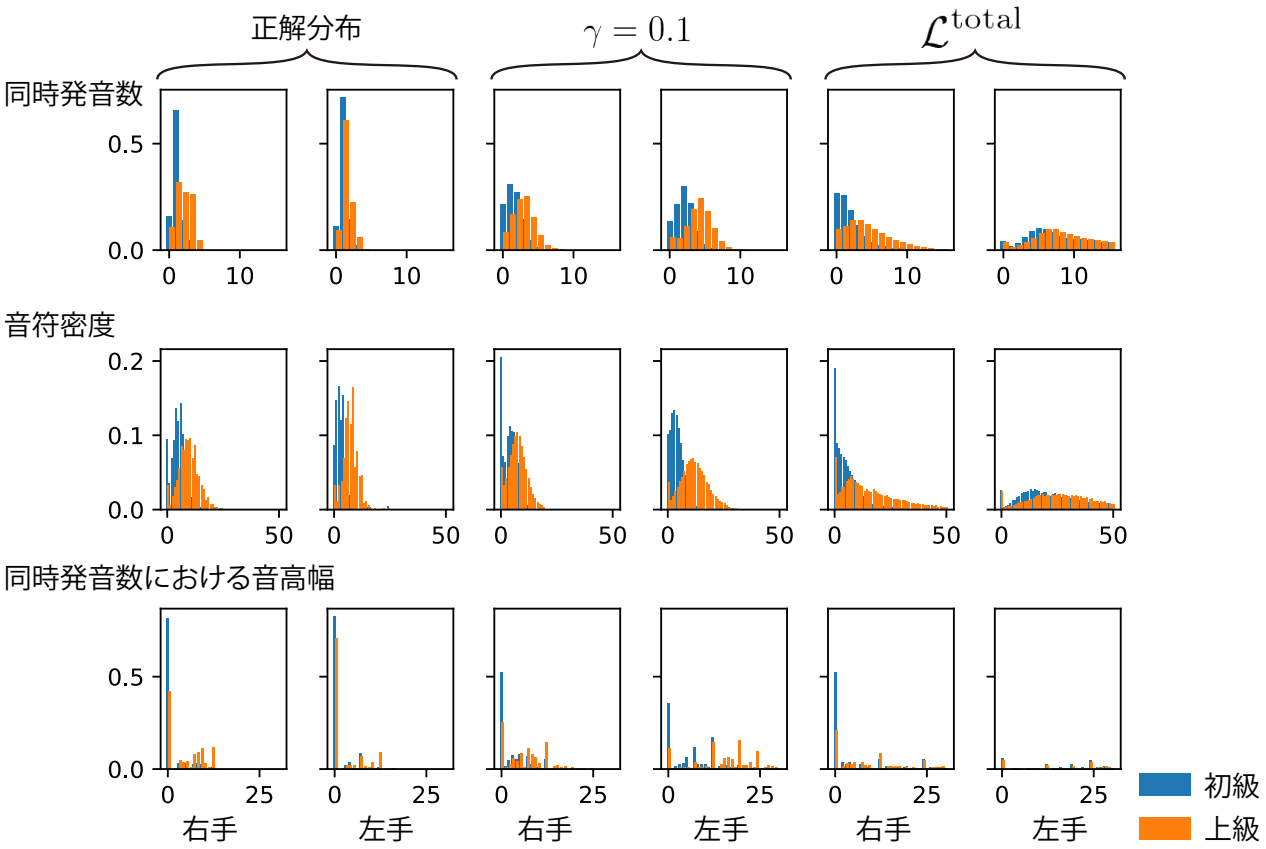

図 4 同時発音数，同時発音における音高幅，音符密度のヒストグラム．見やすさのため横軸は途中までに省略してある．

ここで，ReLU はランプ関数を表す．C_n は必ず 0 以上の整数なので，$-C_n$ は 0 以下になる．$i = k$ の時，$\mathrm{ReLU}(-C_n + k)$ は C_n の要素のうち，k 以上の要素はランプ関数の影響で 0 となる． 反対に $k-1$ だった要素は 1，$k-2$ だった要素は 2，…，0 の要素は k となる．このように考えると f，g，h は（$k-1$ の数）$\times 1+$（$k-2$ の数）$\times 2+\cdots+$（0 の数）$\times k$ を表していることがわかる．また，k と $k-1$ の 2 項間の差は（$k-1$ の数）+（$k-2$ の数）$\cdots+$（0 の数）となる．さらに，次の式のように 2 項間の差の差を取ることで，求める値を導出できる．

$$\begin{aligned} Q_i^{\text{poly}}(\mathbf{Y}) &= (f_{i+2}(\mathbf{Y}) - f_{i+1}(\mathbf{Y})) - (f_{i+1}(\mathbf{Y}) - f_i(\mathbf{Y})) \\ &= f_{i+2}(\mathbf{Y}) - 2f_{i+1}(\mathbf{Y}) + f_i(\mathbf{Y}) \qquad (18) \\ Q_i^{\text{deg}}(\mathbf{Y}) &= (g_{i+2}(\mathbf{Y}) - g_{i+1}(\mathbf{Y})) - (g_{i+1}(\mathbf{Y}) - g_i(\mathbf{Y})) \\ &= g_{i+2}(\mathbf{Y}) - 2g_{i+1}(\mathbf{Y}) + g_i(\mathbf{Y}) \qquad (19) \\ Q_i^{\text{dens}}(\mathbf{Y}) &= (h_{i+2}(\mathbf{Y}) - h_{i+1}(\mathbf{Y})) - (h_{i+1}(\mathbf{Y}) - h_i(\mathbf{Y})) \\ &= h_{i+2}(\mathbf{Y}) - 2h_{i+1}(\mathbf{Y}) + h_i(\mathbf{Y}) \qquad (20) \end{aligned}$$

$Q_i(\mathbf{Y}) =$ は i である要素の個数を示す．また，これらを確率分布とみなすため，足して 1 になるように調整する．

$$T^{\text{poly}} = \sum_{i=0}^{I^{\text{poly}}} Q_i^{\text{poly}}(\mathbf{Y}) \qquad (21)$$

$$T^{\text{deg}} = \sum_{i=0}^{I^{\text{deg}}} Q_i^{\text{deg}}(\mathbf{Y}) \qquad (22)$$

$$T^{\text{dens}} = \sum_{i=0}^{I^{\text{dens}}} Q_i^{\text{dens}}(\mathbf{Y}) \qquad (23)$$

$$\frac{Q_i^{\text{poly}}(\mathbf{Y})}{T^{\text{poly}}},\ \frac{Q_i^{\text{deg}}(\mathbf{Y})}{T^{\text{deg}}},\ \frac{Q_i^{\text{dens}}(\mathbf{Y})}{T^{\text{dens}}} \qquad (24)$$

ただし，I^{poly}，I^{deg}，I^{dens} をそれぞれのヒストグラムの階級数とする．

同様にして，$\hat{C}_n$ を用いて，推定時の分布，$\hat{Q}^{\text{poly}}$，$\hat{Q}^{\text{deg}}$，$\hat{Q}^{\text{dens}}$ を計算する．

ここで，Q と $\hat{Q}$ のヒストグラム同士の距離を測るため JS ダイバージェンス D_{JS} は次のように求められる．

$$M = \frac{(Q + \hat{Q})}{2} \qquad (25)$$

$$D_{KL}(Q \parallel M) = \sum_{i=1}^{I} M_i (\log M_i - \log Q_i) \qquad (26)$$

$$D_{JS}\left(Q \parallel \hat{Q}\right) = \frac{D_{KL}(Q \parallel M) + D_{KL}\left(\hat{Q} \parallel M\right)}{2} \qquad (27)$$

D_{KL} は KL ダイバージェンスを表す．以下のように損失

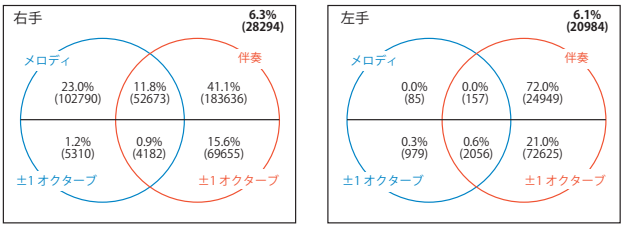

図 5 ピアノ譜に含まれる音符の由来

関数を定義する．

$$\mathcal{L}^{\text{poly}} = D_{\text{JS}}\left(\mathbf{Q}^{\text{poly}} \parallel \hat{\mathbf{Q}}^{\text{poly}}\right) \tag{28}$$

$$\mathcal{L}^{\text{deg}} = D_{\text{JS}}\left(\mathbf{Q}^{\text{deg}} \parallel \hat{\mathbf{Q}}^{\text{deg}}\right) \tag{29}$$

$$\mathcal{L}^{\text{dens}} = D_{\text{JS}}\left(\mathbf{Q}^{\text{dens}} \parallel \hat{\mathbf{Q}}^{\text{dens}}\right) \tag{30}$$

最終的に，β, γ, δ をそれぞれの重みとして，次式で与えられる総合的な損失関数 $\mathcal{L}^{\text{total}}$ を最小化するように U-net を学習する．

$$\mathcal{L}^{\text{total}} = \mathcal{L}^{\text{arran}} + \beta \times \mathcal{L}^{\text{poly}} + \gamma \times \mathcal{L}^{\text{deg}} + \delta \times \mathcal{L}^{\text{dens}} \tag{31}$$

4. 評価実験

本章では，提案手法の有効性を評価するために行った 2 つの実験について報告する．まず，拡張バンド譜から音符を選択，削除してピアノ譜を推定することの妥当性を評価する．次に，収集したポピュラー音楽を用いた実験により，提案手法の性能を検証する．

4.1 実験条件

実験には，YAMAHA ミュージックデータショップから収集したバンド譜とピアノ譜の MIDI データ 184 ペア（初級 85 ペア，上級 99 ペア）を使用し，4 分割交差検証を用いて評価を行なった．U-Net は 4 層の畳み込み層と 4 層の逆畳み込み層で構成される．全ての層で $\mathbf{L}$ をスタックし，バッチ正則化を行った．カーネルサイズは 4，ストライドは 2，パディングは 1 に設定した．過学習を防ぐため，全ての逆畳み込み層でドロップアウト（p = 0.5）を適用した．ネットワークの最適化には Adam（lr = 10^{-4}）を用い，$\alpha = 4$ とした．また，特に表記のない場合は，$\epsilon = 1$ とした．各統計量に関するロスの重みパラメータは，$\mathcal{L}^{\text{total}}$ の場合は，$\beta = 1$，$\gamma = 0.1$，$\delta = 0.1$，$\epsilon = 1$ とした．検証データに対し，$\mathcal{F}$ 値が最大となる閾値を右手左手それぞれで採用した．評価尺度には $\mathcal{F}$ 値を利用し，推定結果の統計的妥当性を評価するため $\mathcal{L}^{\text{poly}}$，$\mathcal{L}^{\text{deg}}$，$\mathcal{L}^{\text{dens}}$ を計算した．

4.2 実験結果

図 5 は，ピアノ譜に含まれる音符の由来を表す．バンド譜由来の音符は右手左手どちらも約 75 %であるが，上下 1 オクターブの音符を含めると約 94 %を占める．これより，バンド譜に含まれる音符及びそれらを上下 1 オクターブシフトした音符の集合から音符を削除することでピアノ譜が得られるという仮定の妥当性を確認できる．

表 1 実験結果

	$\mathcal{F}$		$\mathcal{L}^{\text{poly}}$ ($\times 10^4$)	$\mathcal{L}^{\text{dens}}$ ($\times 10^4$)	$\mathcal{L}^{\text{deg}}$ ($\times 10^4$)
	右手	左手			
$\mathcal{L}^{\text{arran}}$ のみ	**0.56**	0.26	37	0.79	**42**
$\beta = 1$	0.53	0.23	67	1	84
$\gamma = 0.1$	**0.56**	**0.27**	**33**	**0.54**	**42**
$\delta = 0.1$	0.36	0.17	89	4	136
$\mathcal{L}^{\text{total}}$	0.33	0.15	172	9	215
$\mathcal{L}^{\text{total}}, \epsilon = 0.625$	0.50	0.18	224	13	237

表 1 に，U-net を $\mathcal{L}^{\text{arran}}$ で学習した場合と，1 つずつ音符統計量を追加した場合 (例えば，$\beta = 1$ の時は $\gamma = \delta = 0$)，$\mathcal{L}^{\text{total}}$ で学習した場合の評価値を示す．ヒストグラムに関する制約を入れることで，右手の $\mathcal{F}$ 値は $\mathcal{L}^{\text{arran}}$ のみと $\gamma = 0.1$ の場合が最も高く，左手は $\gamma = 0.1$ の場合に最も高くなっている．この理由として，右手に関しては基本的にはバンド譜のメロディパートをそのまま使うことが多いので，統計量に関する損失関数を加えたことで音符単位での学習の収束が遅くなったと考えられる．左手に関しては，バンド譜の伴奏パートは音符数の変動が大きいので，音符密度に関する統計量を加えることで音符単位の学習に貢献したと考えられる．また，$\mathcal{L}^{\text{total}}$ で学習した場合 $\mathcal{F}$ 値は悪化した．これはすべての音符統計量を足した場合のスケールを考慮していなかったためだと考えられる．しかし，同じ条件で $\epsilon = 0.625$ とした場合には改善が見られた．さらに，推定結果の統計的妥当性に関する評価値は $\gamma = 0.1$ とした時に最善であった．その他の場合は $\mathcal{L}^{\text{arran}}$ のみで学習した場合よりも増加した．$\mathcal{L}^{\text{dens}}$ のスケールが相対的に小さいことを考慮すると，まず $\mathcal{L}^{\text{arran}}$ を重視して音符単位での学習が十分に収束した場合に，音符統計量を加えることで，相乗効果として性能向上が期待できる．

図 4 は，正解ピアノ譜，推定されたピアノ譜から得られた難易度別の同時発音数，同時発音における音高幅，小節あたりの音符密度のヒストグラムを表す．ここでは，正解分布と，評価値が最大となる $\gamma = 0.1$ の時の分布と $\mathcal{L}^{\text{total}}$ で学習した分布を表している．$\gamma = 0.1$ の場合では，やはり音符密度に関してのヒストグラムは正解に近い．また，音高幅について見ると，右手に関しては，ほぼ 1 オクターブの範囲に収まっており，理想的な分布になっていると言える．$\mathcal{L}^{\text{total}}$ の分布に関しては，同時発音数や，音符密度の分布が右に寄った分布になっている．

図 6 には，正解ピアノ譜と，$\mathcal{L}^{\text{arran}}$ のみと $\gamma = 0.1$ で学習した推定ピアノ譜の例を示す．右手に関しては，どちら

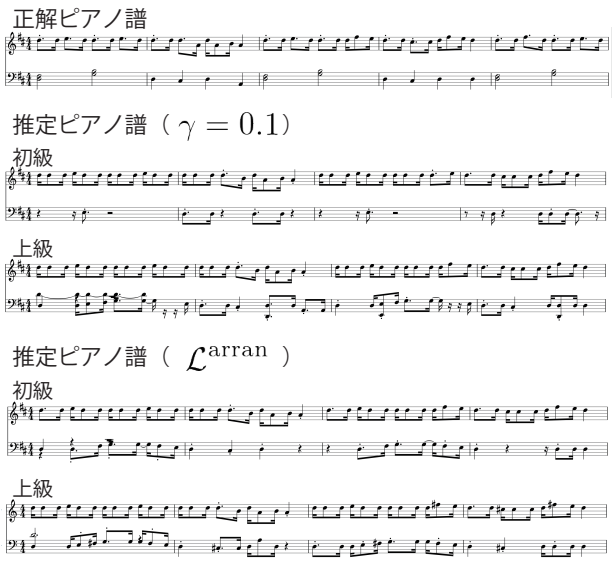

図 6 $\mathcal{L}^{\text{arran}}$ のみと $\gamma = 0.1$ で学習した推定ピアノ譜の例．正解ピアノ譜は初級である．

の場合もほぼ正解ピアノ譜と同じである．$\gamma = 0.1$ で学習した場合は，初級の左手の音符の数が $\mathcal{L}^{\text{arran}}$ のみの場合よりも少なくなっており，正解ピアノ譜と近いと言える．しかし，実際に聞いて見ると，$\mathcal{L}^{\text{arran}}$ のみの場合の方が自然だと感じられたので，今後は，定量的な評価のみでなく，人による主観的な評価も必要である．

5. おわりに

本稿では，深層学習を用いて，ピアノ譜の難易度ごとの統計的性質に基づくピアノ編曲の手法を提案した．今後の課題としては，統計量やオクターブシフトした音符の重みパラメータの調整や，主観評価実験の実施に加え，移調やテンポに代表されるような難易度を定める要素を統計的に調査し，初級・上級という離散的な難易度の枠組みを超えて連続的に難易度を制御することが挙げられる．

謝辞 本研究の一部は，JSPS 科研費 No. 19H04137, No. 20K21813 および JST さきがけ No. JPMJPR20CB の支援を受けた．

参考文献

[1] Chiu, S.-C., Shan, M.-K. and Huang, J.-L.: Automatic system for the arrangement of piano reduction, *Proc. International Symposium on Multimedia*, pp. 459–464 (2009).
[2] Onuma, S. and Hamanaka, M.: Piano Arrangement System Based On Composers' Arrangement Processes, *Proc. International Computer Music Conference*, pp. 191–194 (2010).
[3] Huang, J.-L., Chiu, S.-C. and Shan, M.-K.: Towards an automatic music arrangement framework using score reduction, *ACM Transactions on Multimedia Computing, Communications, and Applications*, Vol. 8, No. 1, pp. 8:1–8:23 (2012).
[4] Takamori, H., Sato, H., Nakatsuka, T. and Morishima, S.: Automatic arranging musical score for piano using important musical elements, *Proc. Sound and Music Computing Conference*, pp. 35–41 (2017).
[5] Nakamura, E. and Yoshii, K.: Statistical Piano Reduction Controlling Performance Difficulty, *APSIPA Transactions on Signal and Information Processing*, No. e13, pp. 1–12 (2018).
[6] Wang, Z., Chen, K., Jiang, J., Zhang, Y., Xu, M., Dai, S., Gu, X. and Xia, G.: POP909: A Pop-Song Dataset for Music Arrangement Generation, *Proc. International Society for Music Information Retrieval*, pp. 38–45 (2020).
[7] Tuohy, D. and Potter, W.: A genetic algorithm for the automatic generation of playable guitar tablature, *Proc. International Computer Music Conference*, pp. 499–502 (2005).
[8] Yoshinaga, Y., Fukayama, S., Kameoka, H. and Sagayama, S.: Automatic arrangement for guitars using hidden Markov model, *Proc. Sound and Music Computing Conference*, pp. 450–456 (2012).
[9] Hori, G., Kameoka, H. and Sagayama, S.: Input-output HMM applied to automatic arrangement for guitars, *Journal of Information Processing*, No. 2, pp. 264–271 (2013).
[10] Maekawa, H., Emura, N., Miura, M. and Yanagida, M.: On machine arrangement for smaller wind-orchestras based on scores for standard wind-orchestras, *Proc. International Conference on Music Perception and Cognition*, pp. 268–273 (2006).
[11] Crestel, L. and Esling, P.: Live orchestral piano, a system for real-time orchestral music generation, *Proc. Sound and Music Computing Conference*, pp. 434–442 (2017).
[12] Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M. and Eck, D.: Music transformer: Generating music with long-term structure, *Proc. International Conference on Learning Representations* (2019).
[13] Nakamura, E. and Sagayama, S.: Automatic piano reduction from ensemble scores based on merged-output hidden Markov model, *Proc. International Computer Music Conference*, pp. 298–305 (2015).
[14] Nakamura, E., Ono, N. and Sagayama, S.: Merged-output HMM for piano fingering of both hands, *Proc. International Society for Music Information Retrieval*, pp. 531–536 (2014).
[15] Yonebayashi, Y., Kameoka, H. and Sagayama, S.: Automatic decision of piano fingering based on a hidden Markov models, *Proc. International Joint Conference on Artificial Intelligence*, pp. 2915–2921 (2007).
[16] Ronneberger, O., Fischer, P. and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241 (2015).