

ボーカル付きポップス楽曲の 音響信号からのピアノ譜自動生成

越井 琢巳^{1,†1,a)} 斎藤 博昭^{1,†1,b)}

概要: 本稿では、ボーカル付きのポップス楽曲の音響信号入力をもとにして、原曲の雰囲気感を再現したピアノ編曲譜を生成する自動『耳コピ』手法を提案する。和音に含まれる単音がそれぞれ発音されるタイミングを2次元で記述したリズムパターン行列 (Rhythm Pattern Matrix, RPM) を定義し、これを用いたリズムパターン自動生成アルゴリズムを考案した。最適ナリズムパターンをコサイン類似度を用いたスペクトログラム比較手法により選出する。ピアノ経験者を被験者としたアンケート調査を行ったところ、楽曲の面白さという評価点において、特に優位性が見られた。

キーワード: 自動編曲, 耳コピ, リズムパターン行列 (RPM), スペクトログラム間類似度

AUTOMATIC ARRANGEMENT FROM WAVEFORMS OF POPULAR VOCAL SONGS TO PIANO SCORES

KOSHII TAKUMI^{1,†1,a)} SAITO HIROAKI^{1,†1,b)}

1. はじめに

本稿では、ボーカル付きのポップス楽曲を『耳コピ』し、ピアノ譜を生成する手法の提案を行う。『耳コピ』とは、耳で聴いた楽曲を楽譜情報なしで再現演奏することである。市販されている楽曲の多くは楽譜という形では売られておらず、CDのように音源のみが販売されている。音源しか手に入らない場合に音響信号情報だけでその楽曲を演奏するのは、ピアノ演奏者にとっても訓練が必要であり、また、ピアノ以外の楽器で演奏された楽曲をピアノ曲風に再現するというテクニックも必要となる。本研究は、こういった『耳コピ』演奏の自動化を目的としている。

2. 研究背景

『耳コピ』のように楽曲の音響信号に対し何らかの形で処理あるいは加工を行うこの分野は、音楽情報処理あるいは Music Information Retrieval(MIR) と呼ばれ、近年盛んに行われている。リズムという時間的秩序や西洋音楽の和声という和音のルールが存在するなどの特徴により音声信号処理とは異なるタスクが多く、例えば音響信号から和音を推定するタスクに関していえば、The International Society for Music Information Retrieval(ISMIR) での MIREX という大会で毎年競争が行われている。また最近の機械学習分野の発展が、音楽情報処理にも大きな影響を与えている。以下に『耳コピ』に関する先行研究の例を挙げる。

2.1 Google の Magenta Project について

Google 社の Magenta Project によって、機械学習を用いた音楽処理の研究が進められている。2017 年に Hawthorne らによって提案された手法 [4] は、ピアノ曲の音響信号から採譜を自動で行う。これは言い換えればピアノ曲を『耳

¹ 情報処理学会

IPSJ, Chiyoda, Tokyo 101-0062, Japan

^{†1} 現在、慶応義塾大学 理工学研究科

Presently with Graduate School of Science and Technology,
Keio University

a) allegro@nak.ics.keio.ac.jp

b) hxs@nak.ics.keio.ac.jp

コピ』する手法であるが、入力ピアノ曲であるため編曲するプロセスが不要という点で本手法と異なる。

2.2 自動編曲に関する先行研究

編曲とは、楽曲の主旋律を残したままリズムや伴奏を変えることによって、異なる楽器で演奏できるようにしたり、異なる雰囲気楽曲に仕上げることを指す。この編曲というタスクを自動でできるようにしようという試みは過去にもなされているが、楽譜情報が入力として得られることを前提としているものがほとんどである。大山らの手法 [9] では、画像の印象に合わせてメロディーと和音を残したままベースとリズムパターンを変化させる。楽曲の雰囲気を画像にマッチしたものに変わるので、原曲とは異なる雰囲気楽曲に仕上がる。

しかしながら、『耳コピ』における編曲となると、原曲の雰囲気を残したまま別楽器で演奏する必要がある。原曲の雰囲気を残す編曲システムとしては、Nakamura らの手法 [8] がある。これは、アンサンブル譜の中から隠れマルコフモデルを使用したアルゴリズムで重要音を選出することで、ピアノソロで演奏可能な楽譜へと音符削減を行うというものである。また、Groves の手法 [3] では教師あり確率的文脈自由文法を用いて、元の旋律の雰囲気を残したまま簡略化する。

本手法の前身として挙げられるのが、伴奏用譜面の自動生成手法 [6] である。これは、入力としてポップス楽曲の音響信号を受け取り、その曲の弾き語り伴奏譜を提示するというシステムである。原曲に最も似た和音とリズムパターンの組合せを選択し、原曲に近い形で伴奏が演奏できる譜面を生成する。このシステムでは、伴奏のリズムパターンはあらかじめ用意した 16 個の中から選択される。この 16 個のリズムパターンは手動で生成したものであり、最小音価は 8 分音符である。手動による最小音価 8 分音符のリズムパターン 16 個では、生成される伴奏が単調であるという問題があった。4 節では、この手法と新たに提案する手法との比較を行う。

2.3 自動楽曲生成に関する先行研究

本手法では、生成される伴奏が単純である問題を解消するため、伴奏部分の自動生成を行う。自動で楽曲を生成するという試みは近年深層学習の分野で特に多くみられ、例えば Yang らの MidiNet [11] や Homma らによる手法 [5] では、敵対的生成ネットワーク (Generative Adversarial Networks, GAN) を用いて旋律を生成する。機械学習による楽曲生成の多くは旋律生成に傾向しており、ピアノの伴奏部の生成を行う研究は多くない。

一方で、Dong らは GAN を用いてドラム・ピアノ・ギターなどの演奏パートを同時に生成し多声楽曲を自動で作曲するという手法 [1] を提案している。この手法では、ピ

アノパートは単音ではなく伴奏という形で出力される。

2.4 本研究の立ち位置

本研究は採譜システムではなく、ボーカルパートのあるピアノ曲ではない音楽音響信号をピアノ曲に編曲してユーザに提示するシステムである。出力がピアノ譜であるにもかかわらず、入力にポップス楽曲の音響信号を取るという点で先行研究と異なる。また、出力はピアノの単旋律ではなく、両手で弾けるソロ演奏向けの楽譜で、右手の譜面は原曲のボーカルパート、左手の譜面は伴奏部分となる。

3. システム構成

本稿で提案する自動『耳コピ』システムについて記述する。本手法の概要を図 1 に示す。それぞれのステップについて詳しく説明する。

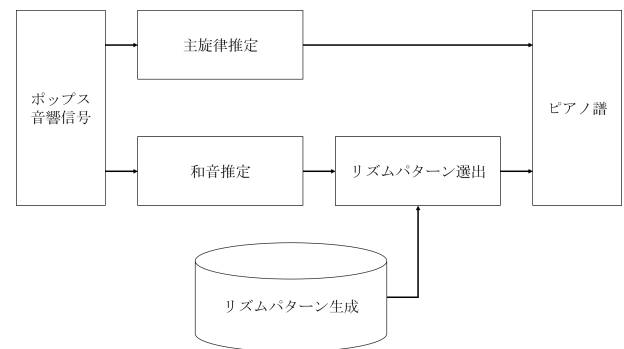


図 1 提案システム概図

3.1 主旋律推定

本手法では原曲のボーカルパートを主旋律とし、それを右手譜として出力する。主旋律抽出には Salamon らの Melodia [10] を用いた。この手法は、音響信号のデータ列のあるまとまりごとに代表となるピッチを与えて代表ピッチの列 (ピッチ外形) を作り、主旋律的ピッチ外形と非主旋律的ピッチ外形に分類することで主旋律成分を分離するというものである。MIREX2011^{*1} の主旋律抽出タスクにおいて最高スコアを達成した。

3.2 和音推定

主旋律推定と並行して、入力された楽曲の和音を推定する。リズムパターン自体は和音情報を持たないため、リズムパターンに当てはめるための和音が必要となるからである (3.3 節で詳しく述べる)。和音推定には Mauch らの Chordino [7] を用いた。MIREX2009^{*2} に登場して以来、MIREX の和音推定部門において達成すべきスコアの

^{*1} https://www.music-ir.org/mirex/wiki/2011:MIREX2011_Results, 閲覧日: 2018 年 11 月 20 日。

^{*2} https://www.music-ir.org/mirex/wiki/2009:Audio_Chord_Detection_Results, 閲覧日: 2018 年 11 月 20 日。

目安として広く用いられている手法である。非負値最小二乗クロマ (Non-Negative Least Squares Chroma, NNLS Chroma) という各ピッチクラス^{*3}の強さを表す特徴量を計算し、隠れマルコフモデルで和音を推定する。本手法では『M』, 『m』, 『M/3』, 『M/5』, 『M₆』, 『m₆』, 『7』, 『M₇』, 『m₇』, 『m₇⁻⁵』, 『dim』, 『aug』の和音の中から推定する。

3.3 リズムパターン生成

本手法では、あらかじめさまざまなリズムパターンを生成し、これら複数通りのリズムパターンに推定和音を当てはめ、原曲と最も似ている伴奏を選択する。ここでは、そのリズムパターン生成手法について説明する。

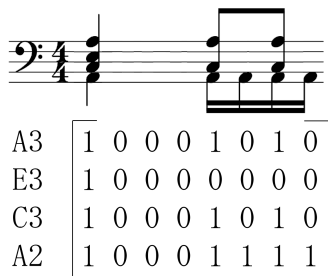


図 2 リズムパターン行列 (RPM)

3.3.1 リズムパターン行列 (Rhythm Pattern Matrix, RPM) の定義

タイムステップ $t \in T$ で発音開始のイベント (Note-on Event) があれば 1, なければ 0 とする行列 $M \in \mathbb{Z}^{N \times T}$ を RPM と定義する。ただし N は和音を持つ構成音の数とし、本システムにおいては $N = 4$ とする。和音構成音数が 3 のものは根音を 1 オクターブ上げた音を加えて 4 とし、和音構成音が 5 以上のものは根音から 4 個目までの音を使用する。また、 T はリズムパターンの時間長である。1 つの RPM に対し 1 つの和音を与えることになるため、和音が切り替わるまでのタイムステップ数といえる。この T に与える値を変化させることで 3 拍子や 5 拍子の曲にも柔軟に対応させることができる。例えば、3 拍ごとに和音が切り替わり、最小音価が 8 分音符である伴奏を作るならば $T = 2[\text{steps per beat}] \times 3[\text{beats}] = 6[\text{steps}]$ とする。

図 2 は、和音『Am』のピアノ譜を RPM で表現する例である。和音が変われば和音構成音も変化するため、RPM 自体は和音情報を持たない^{*4}。また逆に、RPM に和音構成音情報を与えることで楽譜となる。

RPM 最大の特徴は、発音のあったタイムステップのみが示され、音価は分からないことである。図 2 に関して言えば、有音区間であるにもかかわらず、値が 0 になってい

^{*3} ピッチクラス: 同じ音名を与えられている音の集合。

^{*4} RPM は和音構成音中の何番目の音が発音されるかという情報だけを持ち、その音が具体的にどのキーであるかは和音情報を与えられるまで決定されない。

る部分があるなど、直感にはそぐわない部分が目立つ。これは、ピアノという楽器でポップス楽曲の伴奏を演奏する場合、原則的には和音ごとにサスティンペダルを踏みかえて演奏するという前提に立ったモデリングであるためである。サスティンペダルを踏んでいる間は音が伸び続けるため、有音区間の終わるタイミングは和音が切り替わる時、すなわち次の RPM に移るときである。

3.3.2 RPM の分離

ピアノ伴奏譜の特徴を踏まえ楽譜が密になりすぎないように、RPM をリズム成分 $r \in \mathbb{Z}^T$ とボイシング成分 $V \in \mathbb{Z}^{N \times K}$ に分けてから (図 3)、次列生成を行う。ただし K とは、リズム成分中に含まれる 1 の個数である。ここでリズム成分は、各 t における発音イベントの有無を表し、ボイシング成分はその発音イベントで発音された和音構成音を表している。

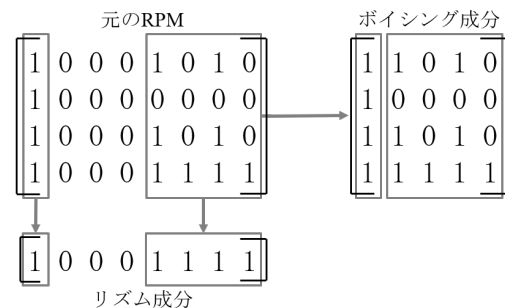


図 3 RPM の分離

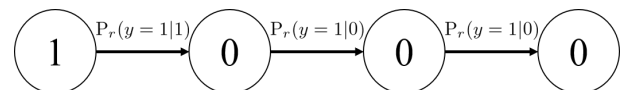


図 4 リズム成分の次列生成

3.3.3 RPM の次列生成

RPM の各成分に対して、前列の状態から次列を生成する。まず、あらかじめ用意した楽曲データの伴奏部分からリズム成分 r を抽出し、前ステップが $x (x = 0 \text{ or } 1)$ のときの次ステップ y の値が 0 となる個数 $\text{Number}_r(y = 0|x)$, 1 となる個数 $\text{Number}_r(y = 1|x)$ をそれぞれ計算する。それをもとに、前ステップが x であるときに $y = 1$ となる確率 $P_r(y = 1|x)$ を求める [式 (1)]。

$$P_r(y = 1|x) = \frac{\text{Number}_r(y = 1|x)}{\sum_{bool=0}^1 \text{Number}_r(y = bool|x)} \quad (1)$$

前の要素が得られたとき、上の確率をもとに次の要素を決定する (図 4)。[0, 1) の範囲の乱数 r を取得し、例えば前の要素が 1 のときは、 $r < P_r(y = 1|1)$ であれば 1 を、そうでなければ 0 を生成する。

次に、ボイシング成分でも同様に、前列の状態から、次

の列のそれぞれの要素について 1 となる確率を求める．ボイシング成分を構成する列 $\mathbf{v}_k^T (1 \leq k \leq K)$ は、格納されている値が 2 値であるため、それぞれ 2^N 通りの状態を持つといえる．そこで、状態番号 s を $0 \leq s \leq 2^N - 1$ の整数とし、各列 $\mathbf{v}_k^T$ の状態番号 $S(\mathbf{v}_k)$ を以下の [式 (2)] のように定義する．

$$S(\mathbf{v}_k) = [2^0, 2^1, \dots, 2^n, \dots, 2^{N-1}] \mathbf{v}_k^T \quad (2)$$

あらかじめ用意した楽曲データの伴奏部分からボイシング成分 $\mathbf{V}$ を抽出し、前列の状態番号が s であるときの次列 $\mathbf{y}^T \in \mathbb{Z}^{N \times 1}$ の n 行目の要素が 1 である個数 $\text{Number}_V(y_n = 1|s)$ を求める．これをもとに確率 $P_v(y_n = 1|s)$ を計算する [式 (3)]．

$$P_v(y_n = 1|s) = \frac{\text{Number}_V(y_n = 1|s)}{\sum_{bool=0}^1 \text{Number}_V(y_n = bool|s)} \quad (3)$$

このようにすることで、新たに i 番目の列の状態番号 $S(\mathbf{v}_i)$ が与えられたとき、次列 $\mathbf{v}_{i+1}^T$ の各要素について 1 となる確率 $P(v_{n,i+1} = 1)$ は、 $P_v(y_n = 1|s)$ を用いて、[式 (4)] のように表せる．

$$P(v_{n,i+1} = 1) = P_v(y_n = 1|S(\mathbf{v}_i)) \quad (4)$$

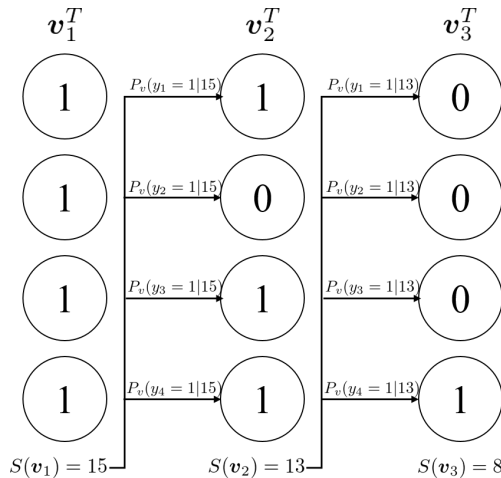


図 5 ボイシング成分の次列生成

これを用いて、リズム成分同様に次列を生成する (図 5)．

リズム成分について、これらの確率をもとに次列生成を行い、新たな RPM のリズム成分 $\mathbf{M}_r$ を生成する．続いて、生成した $\mathbf{M}_r$ 中の 1 の数 K を計算し、ボイシング成分を生成する．

1 つ前の状態を見るだけで伴奏が生成できるのは、使用するキーを和音構成音のみとし不協和音を避けているためである．また、Note-on に着目した行列で同音連打表現を可能にしながらも、リズム成分とボイシング成分への分離によって同音連打の頻発を防いでいることも、自然な伴奏になる理由の一つとして挙げられる．

3.4 リズムパターン選出

伴奏の選択にはスペクトログラム間類似度を用いた．スペクトログラムは周波数 $F \times$ 時間 T の行列と見なせる．まず、比較する二つのスペクトログラムの絶対値 $\mathbf{A} \in \mathbb{R}^{F \times T}$ と $\mathbf{B} \in \mathbb{R}^{F \times T}$ に対し、周波数方向のコサイン類似度時間平均 $\text{Sim}_f(\mathbf{A}, \mathbf{B})$ と時間方向のコサイン類似度周波数平均 $\text{Sim}_t(\mathbf{A}, \mathbf{B})$ を求める [式 (5)]．

$$\begin{aligned} \text{Sim}_f(\mathbf{A}, \mathbf{B}) &= \frac{1}{T} \sum_{t=1}^T \frac{\mathbf{a}_t \cdot \mathbf{b}_t}{|\mathbf{a}_t| |\mathbf{b}_t|}, \\ \text{Sim}_t(\mathbf{A}, \mathbf{B}) &= \frac{1}{F} \sum_{f=1}^F \frac{\mathbf{a}_f \cdot \mathbf{b}_f}{|\mathbf{a}_f| |\mathbf{b}_f|} \end{aligned} \quad (5)$$

ただしベクトル $\mathbf{a}, \mathbf{b}$ はそれぞれ行列 $\mathbf{A}, \mathbf{B}$ の成分である． $\text{Sim}_f(\mathbf{A}, \mathbf{B})$, $\text{Sim}_t(\mathbf{A}, \mathbf{B})$ のそれぞれを正規化した和をスペクトログラム間類似度 $\text{Sim}(\mathbf{A}, \mathbf{B})$ とした．

生成された複数のリズムパターンに対し推定和音を適用させ、原曲とのスペクトログラム間類似度が最も高くなるリズムパターンを選択する．窓幅や周波数領域を調整すれば、動的時間伸縮法 (Dynamic Time Warping, DTW) による類似度計算よりも比較的高速に実行でき、異なるサンプリングレート間での信号比較も可能である．



図 6 出力例

3.5 出力例

出力は標準 MIDI ファイル (Standard MIDI File, SMF) である．本システムの入出力例を SMF から楽譜に変換して表示すると、図 6 のようになる．これは、RWC 研究用音楽データベース [2] のポピュラー音楽データベースから Track7 のサビ部分を抜粋し音響信号データと拍子、BPM を入力した際の出力結果である．右手の譜面は原曲のボーカルの部分を演奏するものである．左手の譜面は、このシ

ステムが生成する伴奏の部分であり、原曲に最も似ている和音とリズムパターンの組合せを出力している。

4. 実験と評価

本システムにポップス音楽を入力し、まずは、メロディー推定と和音推定の部分の評価を行った。これは、本手法の主要な部分となるリズムパターン選出より以前の、前処理の段階で、評価に悪い影響を与える結果が出ていないかを確認するためである。次に、得られた楽譜に対して主観評価実験を行った。

4.1 実験環境

入力には、RWC 研究用音楽データベース [2] のポピュラー音楽データベースから初めの 10 曲を選択し、サビ部分の WAV データを使用した。RPM 生成のためのピアノ楽曲には、同じく RWC 音楽データベース [2] のクラシック音楽データベースから、ピアノトラックが含まれる 26 曲を選択し、ピアノトラック 11,514 拍分の SMF データを使用した。また、和音情報と RPM からスペクトログラムを得るため、Roland 製電子ピアノ『HP603』の『GM2 ピアノ 1』音源を単音ごとに録音して使用した。4.3 節の主観評価実験では、生成された SMF ファイルを、同じく『HP603』の SMF 再生機能を用いて 3 名のピアノ経験者に聴いてもらい、アンケート評価実験を行った。

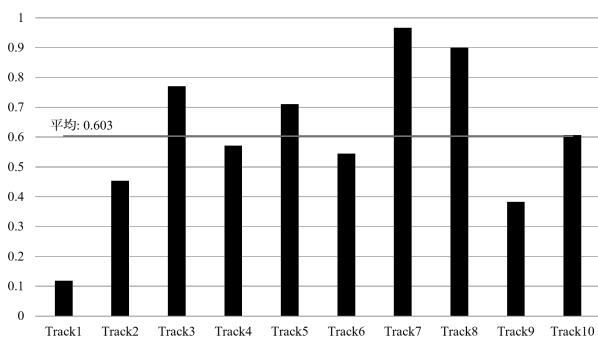


図 7 メロディー推定の適合率

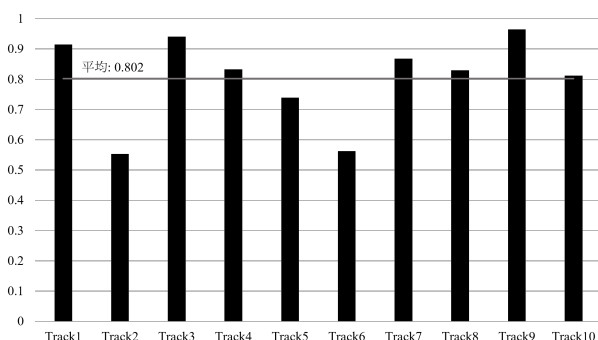


図 8 和音構成音の F 値

4.2 メロディー推定の適合率と和音構成音の F 値

メロディー推定の適合率は入力楽曲ごとにばらつきがあった (図 7)。

特に平均値 0.603 を超えなかった Track1, 2, 4, 6, 9 に関しては、メロディー推定の適合率の低さが主観評価実験の結果に悪い影響を与えると予想される。次に、推定された和音と、原曲の和音との間で、和音が構成する音の F 値を図 8 に示す。

F 値の値が平均値 0.802 を超えなかった Track2, 5, 6 が主観評価実験の結果に悪い影響を与えると予想される。4.3 節ではこれらの結果を踏まえた比較実験を行う。

4.3 主観評価実験

出力譜が面白いかどうか、原曲に似ているかどうか、そして、実際にシステムを利用したいと感じるかの 3 点で 5 段階評価を行った。比較対象としては、2 拍子・4 拍子に対応した 16 個のリズムパターンをあらかじめ用意する手法 [6] を選択した。主観評価実験の結果は図 9 のようになった。

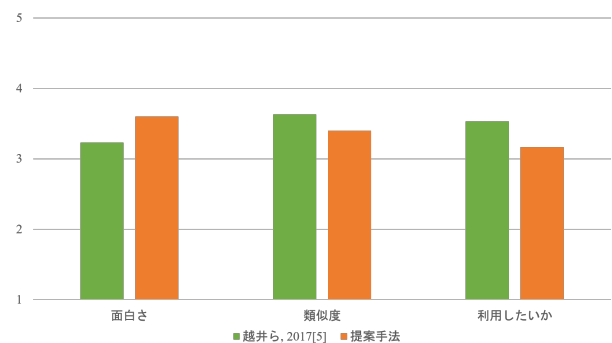


図 9 全 Track での評価スコア平均

[6] の手法では最小音価が 8 分音符のリズムパターン 16 個の中から選択される。このため、最も細かい音でも 8 分音符という簡単な楽譜という印象になる。全部で 16 個と数も多くないためワンパターンにもなりうる。一方で本手法は、最小音価 16 分音符のリズムパターン 64 個を用いてさらに複雑な伴奏を生成し、楽曲の面白さという評価点で高いスコアを達成した。この 64 個のリズムパターンはピアノ楽曲データをもとに自動生成しており、演奏不可能な同音連打などは発生しにくくなっている。しかしながら、伴奏が複雑化したことで類似度のスコアは下がってしまう結果となった。また、利用したいかという点でもスコアが低く、[6] の手法はリズムパターンが多少単純ではあるものの伴奏としては無難であるという点で評価される結果となった。

次に、4.2 節において前処理に問題がなかったと判断された Track3, 7, 8, 10 での評価平均を図 10 に示す。それとは逆に、図 11 は前処理の結果が悪い影響を与えると予想

される Track1, 2, 4, 5, 6, 9 の評価平均である。

図 10 から示される通り、Track3, 7, 8, 10 は前処理の評価が良いために、全体的にスコアが高くなっている。[6] の手法は前処理の影響を強く受けており、さらにスコアが高くなっている。次に図 11 についてであるが、Track1, 2, 4, 5, 6, 9 でのスコアは全体的に下がってはいるものの、[6] の手法に比べ本手法は、前処理からの悪影響を受けにくいという結果となった。[6] の手法は伴奏の音価が全体的に長いため、和音が間違っていると不協和音として響きやすい。しかしながら、本手法は音価が全体的に短いため、和音が多少間違っても違和感を与えにくい伴奏譜となっている。メロディー推定の適合率・和音構成音の F 値の低さによる違和感を複雑な伴奏でカバーし、面白い楽譜に仕上げる事ができている。

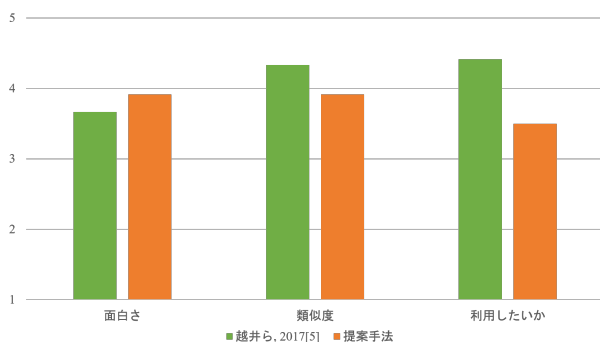


図 10 Track3, 7, 8, 10 での評価スコア平均

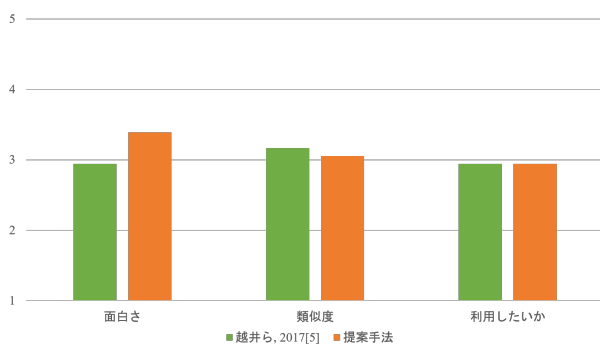


図 11 Track1, 2, 4, 5, 6, 9 での評価スコア平均

5. 結論と今後の展望

楽曲の音響信号データの入力で、ピアノ譜を出力する『耳コピ』システムの提案を行った。楽曲の面白さという点で先行研究の評価を上回った。また、メロディー推定や和音推定といった前処理の部分での性能が低い場合でも、良い評価が得られる伴奏を生成することができた。今後の展望としては、RPM 生成前に明るいまたは暗いなど楽曲の雰囲気や推定し、2.3 節で述べた GAN のような深層学習を用いて雰囲気に適した RPM を動的に生成することで、より原曲の雰囲気に近づけた編曲を行う手法を考えている。

6. 謝辞

本稿を作成するにあたって実験に使用した RWC 研究用音楽データベース [2] の作成者の方々および、電子ピアノ『HP603』の Roland 社、また、実験に協力してくださった方々に対し感謝申し上げます。

参考文献

- [1] Dong, H.-W. and Yang, Y.-H.: Convolutional Generative Adversarial Networks with Binary Neurons for Polyphonic Music Generation, *ISMIR* (2018).
- [2] 後藤真孝, 橋口博樹, 西村拓一, 岡 隆一: RWC 研究用音楽データベース: 研究目的で利用可能な著作権処理済み楽曲・楽器音データベース, 情報処理学会論文誌, Vol. 45, No. 3, pp. 728–738 (2004).
- [3] Groves, R.: Automatic Melodic Reduction Using a Supervised Probabilistic Context-Free Grammar., *ISMIR*, pp. 775–781 (2016).
- [4] 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, *arXiv preprint arXiv:1710.11153* (2017).
- [5] Homma, N. and Saito, H.: Automatic Image-to-melody Composition Using Deep Learning, *ISMIR* (2018).
- [6] 越井琢巳, 斎藤博昭: スペクトログラム間の類似度に基づくピアノ弾き語り用伴奏譜の自動生成, 研究報告音楽情報科学 (MUS), Vol. 2017, No. 49, pp. 1–6 (2017).
- [7] Mauch, M. and Dixon, S.: Approximate Note Transcription for the Improved Identification of Difficult Chords., *ISMIR*, pp. 135–140 (2010).
- [8] Nakamura, E. and Sagayama, S.: Automatic Piano Reduction from Ensemble Scores Based on Merged-Output Hidden Markov Model, *ICMC* (2015).
- [9] 大山喜冴, 伊藤貴之: DIVA: 画像の印象に合わせた音楽自動アレンジの一手法の提案, 芸術科学会論文誌, Vol. 6, No. 3, pp. 126–135 (2007).
- [10] Salamon, J. and Gómez, E.: Melody extraction from polyphonic music signals using pitch contour characteristics, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 6, pp. 1759–1770 (2012).
- [11] Yang, L.-C., Chou, S.-Y. and Yang, Y.-H.: MidiNet: A convolutional generative adversarial network for symbolic-domain music generation, *arXiv preprint arXiv:1703.10847* (2017).