

# ユーザ好みのスキヤットを強化学習する 初音ミクとのジャムセッションシステム

鶴田 穰士<sup>1,a)</sup> 岡 夏樹<sup>1</sup> 田中 一晶<sup>1</sup>

**概要:** ジャムセッションではボーカルがスキヤットと呼ばれる歌唱法を用いることがある。本研究では、ユーザの演奏に対してシステムが初音ミクを用いてスキヤットを返すという形態のジャムセッションシステムを提案する。さらに、ジャムセッションを通してユーザが好むスキヤットの言葉を強化学習していくことで、ユーザに合ったジャムセッションをすることが可能となる。提案システムの実装を行い、動作を確かめた。

## 1. はじめに

音楽を用いたコミュニケーション手段として、ジャムセッションというものがある。これは楽譜に書かれた通りに演奏するアンサンブルとは異なり、演奏者同士が即興的に演奏していく形態のことである。ある演奏者による演奏が他の演奏者に影響を与え、新たな演奏として返ってくるといったような演奏者間のインタラクションが顕著である点がアンサンブルとの大きな違いである。

しかし、ジャムセッションを行うには他の演奏者の協力が必要である。それを個人でいつでも可能としたのが、ユーザと機械とのジャムセッションシステムであり、これまで多く研究されてきた。演奏形態として、主にピアノ同士のジャムセッションを扱ったものが多い。コード進行をあらかじめ決めておくもの [1] [2] や、ユーザに合わせてコード進行を変えるもの [3]、ユーザがメロディを弾き、それに沿ったコード進行でシステムが伴奏を行い、そのコード進行に応じてユーザがメロディを弾くといったもの [4] がある。しかしこれらのシステムはコード進行といった専門知識を有するジャムセッション熟練者のためのシステムであり、初心者には敷居が高く感じられる。コード進行を有さない打楽器であるドラム同士のジャムセッションシステムの開発もされている [5]。ピアノ同士でコード進行の知識が無い初心者でも気軽に始められるジャムセッションシステムとして Yotam らは「A.I. Duet」を開発した [6]。多数の既存の楽曲のメロディパターンを LSTM を応用し

たニューラルネットワークを用いて学習させ、ユーザが入力したメロディに自然に繋がるようなメロディをシステムが返すというものである。web ブラウザ上で、何の音楽知識がない人でも簡単にジャムセッションを体験できるという点において他のジャムセッションシステムより優れている。

本研究では、先行研究では用いられていないスキヤットを扱ったジャムセッションシステムを提案する。スキヤットとは、ジャムセッションにおいてボーカルが歌声を楽器のように使う歌唱法である。あらかじめ決められた歌詞ではなく、「ラララ〜」や「シャバドゥビ」、「ダバダバ」といった意味の無い音を即興的に歌うことで、ジャムセッションを行う。スキヤットに使われる言葉は、よく使われるものはあるが特に決まりは無く、歌唱者が好きなように歌うものであって、聴き手側でも好みがかたかれやすい。したがって、本研究ではさらに、機械である点を生かし、強化学習を用いて、ユーザとのジャムセッションを通して各々のユーザの好みに合うようなスキヤットの言葉を学習するジャムセッションシステムを提案する。提案システムでは、セッション相手としてスキヤットを歌う初音ミク<sup>\*1</sup>を用いた。初音ミクという一般的に認知されている既存のキャラクターを使うことで先行研究のような楽器が鳴るだけのジャムセッションシステムに比べて、人間の代替としてのリアリティを感じられ、セッション相手の存在をより強く意識することができると考えられる。また、人間には歌いにくい、歌えないような歌を自由に発声可能な点が歌声合成技術の長所である。したがって、ユーザの好みに応じた、人間のスキヤットの型にとらわれない、自由で新しいスタイ

<sup>1</sup> 京都工芸繊維大学 大学院工芸科学研究科  
Graduate School of Science and Technology, Kyoto Institute  
of Technology, Kyoto 606-8585, Japan

<sup>a)</sup> tsurada@ii.is.kit.ac.jp

<sup>\*1</sup> クリプトン・フューチャー・メディア社製の歌声合成ソフトの製品名およびそのパッケージに描かれたキャラクターの名称。

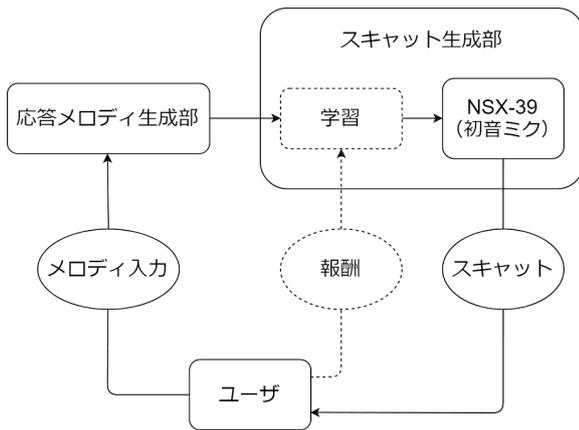


図 1 提案システムの概要図



図 2 提案システムを使用している様子

ルのスカットの生成も期待できると考えられる。

## 2. 提案システム

### 2.1 システム概要

提案システムの概要図を図 1 に示す。本研究では、図中の点線部を新たに作成した。提案システムは大きく分けて、応答メロディ生成部とスカット生成部に区分される。提案システムの大まかな流れは以下の通りである。

- (1) ユーザがキーボードによってメロディを入力する。
- (2) 入力されたメロディをもとに、応答メロディ生成部にて、スカットのメロディを生成する。
- (3) スカット生成部にて、生成されたメロディと学習経験をもとに初音ミクがスカットとして歌う。
- (4) ユーザはそのスカットを自分の好みかどうかで評価して、報酬を与える。
- (5) 与えられた報酬に応じてユーザ好みのスカットを学習し、1 へ戻る。

提案システムを使用している様子を図 2 に示す。ぬいぐるみがある位置から初音ミクのスカット音が流れる。本研究では主にスカット生成部について行ってきたため、応答メロディ生成部については 2.2 節で簡単に紹介し、スカット生成部は 2.3 節にて詳しく説明する。

### 2.2 応答メロディ生成部

応答メロディ生成部では、初心者でも気軽にジャムセッションを行うことが可能である先行研究の「A.I. Duet」[6]を利用した。また、「A.I. Duet」は Magenta\*2 というオープンソースプロジェクトに基づくもので、誰でもシステム改良を行うことが可能となっているため、本研究ではこのシステムを利用して開発を行った。ユーザがキーボードによって、適当なメロディを弾くと、ピアノ音がスピーカーから流れ、それと同時に MIDI 信号が PC に送られる。それをもとに、図 2 のようにディスプレイ上にどの音階をどのタイミングで弾いたのかがリアルタイムで視覚的にわかるように表示される。弾いた音階を表すオブジェクトが時間と共に上へ流れていく仕組みである。赤紫色のオブジェクトがユーザが入力したメロディ、緑色のオブジェクトが出力された応答メロディを表す。

ユーザが入力した MIDI 信号（音階と音の長さ情報のみを扱う）を事前学習したニューラルネットワークを通して、入力したメロディに自然に続くような MIDI 信号を生成し、スカット生成部に送る。ここで使うニューラルネットワークは、Lookback RNN である。Lookback RNN とは、LSTM をもとにした機械学習による自動作曲のためのニューラルネットワークモデルである。楽曲特有の繰り返し構造を学習しやすくしたモデルであり、相手のフレーズを模倣したようなフレーズを返すことが多いジャムセッションシステムに適している。提案システムでは、Magenta による学習済み配布モデルを使用した。

### 2.3 スカット生成部

#### 2.3.1 スカット生成部の概要

スカットには、メロディと言葉が必要である。提案システムでは、メロディは応答メロディ生成部で生成された MIDI 信号をそのまま用いて、言葉のほうを強化学習する。スカット生成には、リアルタイムに初音ミクの歌唱が可能なデバイス NSX-39 を用いた。また、強化学習の手法としては Q 学習を用いた。スカット生成部ではまず、応答メロディ生成部で生成した MIDI 信号情報と、強化学習によって決まる言葉の情報を NSX-39 へ送り、初音ミクがスカットを歌う。次に、ユーザはそのスカットを自分の好みかどうかの 2 択で評価し、図 2 中下部のように PC キーボードを用いて、2 択の評価のそれぞれに割り当てられたキーを 1 つだけ押すことによって報酬を与え、次の演奏入力を開始する。

以上を繰り返すことでユーザ好みのスカットの言葉を強化学習していく。

\*2 <http://magenta.tensorflow.org/>

## 2.4 学習方法

提案システムでは、強化学習手法である Q 学習を用いた。

### 2.4.1 行動

スキヤットに使う言葉の種類を行動とした。以下の点を考慮する。

- 新しいスタイルのスキヤットの生成を期待しているため、一般的にスキヤットに使われる言葉と使われない言葉を含める。「ラ」、「ドゥビ」と「ケ」
- 言葉の繋がりが好みに影響すると考えられるため、1音だけでも成り立つ言葉と2音セットで使われる言葉を含める。「ラ」と「ドゥビ」。
- 言葉間の違いが聴き取りやすいように言葉の種類を選ぶ(異なる母音など)。
- 行動は多すぎると学習に時間が掛かる。

以上のことを踏まえて検討した結果、「ラ」、「ドゥ」、「ビ」、「ケ」の4つの言葉を選択した。これらの内どの言葉を使うかを学習する。

### 2.4.2 状態

今回扱う MIDI 信号は音階と、音の長さ情報のみで構成される。本来メロディは系列データとして扱うべきであるが、提案システムは実時間学習であり、系列データとして扱うと状態数が非常に多くなり、学習に時間が掛かるため、適さないと考えられる。そのためスキヤット1音に対して以下のように離散化したものを状態とした。

- 入力できる音階の中心を境に低いか高いかの2状態。
- 音の長さが0.25秒より短いか長いかの2状態。

さらに、状態数の関係上、系列データとして扱うことはできなかったが、系列データの性質に近づけるため、1つ前の行動を状態として追加した。スキヤットでは音の高さや長さの系列より、言葉の繋がりがより重要であると考えられるからである。行動は4つあるが、1音目は前の行動が無いので、行動無しという状態も追加した5状態となる。

したがって、合計で  $2 \times 2 \times 5 = 20$  状態となる。

### 2.4.3 方策

行動選択確率  $\pi(s, a)$  には、式(1)に示したソフトマックス手法を用いる。

$$\pi(s, a) = \frac{\exp(Q(s, a)/\tau)}{\sum_{p \in A} \exp(Q(s, p)/\tau)} \quad (1)$$

ただし、 $\tau (> 0)$  は温度と呼ばれるパラメータである。温度  $\tau = 0.3$  とした。

### 2.4.4 報酬

ユーザは初音ミクのスキヤット(1フレーズ)が終わるとそのフレーズ全体に対する評価(メロディーでなく言葉に対する評価)を行い、PCキーボードを用いて、報酬を与えることができる。好みであると判断した場合は+1、好みでないと判断した場合は-1の報酬を与え、そのどちらでもない判断した場合はユーザは何もする必要は無く、

表 1 Q 値表

| 状態   |      | 行動   |               |               |               |               |
|------|------|------|---------------|---------------|---------------|---------------|
| 音の高さ | 音の長さ | 前の行動 | 「ラ」           | 「ドゥ」          | 「ビ」           | 「ケ」           |
| 低い   | 短い   | 「ラ」  | 0.197         | 0.143         | -0.276        | -0.230        |
|      |      | 「ドゥ」 | -0.163        | -0.006        | -0.340        | -0.074        |
|      |      | 「ビ」  | -0.264        | -0.257        | 0.049         | -0.360        |
|      |      | 「ケ」  | -0.278        | 0.004         | -0.188        | -0.344        |
|      |      | 無し   | 0.100         | 0.271         | -0.264        | -0.271        |
|      | 長い   | 「ラ」  | -0.337        | 0.645         | -0.243        | -0.107        |
|      |      | 「ドゥ」 | 0.515         | 0.253         | 0.486         | <b>-0.570</b> |
|      |      | 「ビ」  | 0.122         | 0.046         | -0.386        | -0.293        |
|      |      | 「ケ」  | <b>-0.506</b> | -0.263        | -0.344        | <b>-0.522</b> |
|      |      | 無し   | 0.202         | -0.144        | -0.379        | -0.383        |
| 高い   | 短い   | 「ラ」  | -0.385        | 0.044         | -0.332        | -0.190        |
|      |      | 「ドゥ」 | -0.119        | 0.292         | 0.131         | -0.143        |
|      |      | 「ビ」  | -0.210        | -0.092        | -0.265        | -0.247        |
|      |      | 「ケ」  | <b>-0.613</b> | <b>-0.453</b> | -0.410        | <b>-0.556</b> |
|      |      | 無し   | 0.018         | -0.116        | -0.048        | -0.344        |
|      | 長い   | 「ラ」  | -0.340        | 0.004         | -0.210        | -0.410        |
|      |      | 「ドゥ」 | -0.071        | 0.213         | 0.678         | -0.421        |
|      |      | 「ビ」  | -0.350        | 0.516         | -0.147        | -0.415        |
|      |      | 「ケ」  | -0.344        | -0.126        | <b>-0.613</b> | <b>-0.522</b> |
|      |      | 無し   | 0.399         | 0.501         | <b>-0.445</b> | <b>-0.530</b> |

その場合報酬は与えられない。

### 2.4.5 Q 値更新

報酬が与えられたときに、対象となるスキヤットを構成する一音一音に対して更新式(2)によってQ値の更新を行う

$$Q(s, a) \leftarrow Q(s, a) + \alpha(r - Q(s, a)) \quad (2)$$

ただし、学習率  $\alpha = 0.1$  とした。また、Q値の初期値は全て0とした。

## 3. 動作試験

### 3.1 試験環境

著者によるシステムの動作確認を行った。メロディ入力スキヤット出力までを1エピソードとして、250エピソード行った。

### 3.2 結果

250エピソード後のQ値を表1に示す。数値の高い上位10項の背景を薄く塗り、下位10項の文字を太字斜体で表した。また、表2では、言葉の遷移に着目し、表中の数値はそれぞれに対応するQ値の平均を取ったものである。数値の高い上位3項の背景を薄く塗り、下位3項の文字を太字斜体で表した。学習していないシステムと比べて、250エピソード学習を行ったシステムの方が、より自分好みのスキヤットが出力されるように感じられた。

## 4. 考察

表1をみると、最もQ値が高くなったのは、「ドゥ」のあとに高く長い「ビ」が続くスキヤットとなっている。これは一般的によく使われるスキヤット「ドゥビ」と言葉は

表 2 言葉の遷移に着目した Q 値の平均値表

| 前の行動 \ 行動 | 「ラ」           | 「ドゥ」   | 「ビ」           | 「ケ」           |
|-----------|---------------|--------|---------------|---------------|
| 「ラ」       | -0.216        | 0.209  | -0.265        | -0.234        |
| 「ドゥ」      | 0.040         | 0.188  | 0.239         | -0.302        |
| 「ビ」       | -0.175        | 0.053  | -0.187        | -0.329        |
| 「ケ」       | <b>-0.435</b> | -0.210 | <b>-0.389</b> | <b>-0.486</b> |
| 無し        | 0.180         | 0.128  | -0.284        | -0.382        |

同じである。しかし、一般的なスキヤットでは「ドゥビー」と語尾を伸ばすことはなく、短い「ビ」で終わるか、すぐにまた「ドゥビドゥビ」と繰り返すことが多い。これは、提案システムによる新しいスタイルとまではいかないが、一般的に使われないようなスキヤットが生成ができたと考えられるかもしれない。1音目の高く長い音に対しては「ラ、ドゥ」と「ビ、ケ」に顕著な差が表れている。したがって、1音目の高く長いメロディは好みを判断する上で重要な要因となっていることがわかる。また、Q 値の低いところに着目すると、下位 10 項目中 9 つが「ケ」に関連していることがわかる。「ケ」は一般的なスキヤットには使われない言葉であり、歌いにくい言葉ではあるが、聴き取る側も好みになりにくい言葉となる可能性があるといえる。

表 2 をみると、「ドゥ」が総じて 1 番好まれた言葉だったことがわかる。1 音目から数値の高い言葉を遷移順に並べると、「ラドゥビドゥビドゥビ…」と続く。「ドゥ」→「ドゥ」の遷移もしやすいので、たまにドゥが連続するような「ラドゥドゥビドゥビドゥドゥビ…」といったようなスキヤットが好みであったと考えられる。

## 5. 課題点と改善策

### 学習に時間がかかる点

今回の実験では、実時間で 30 分もかかった。改善策として、状態や行動の分け方、パラメータの調整といった学習手法の見直しなどが考えられる。

### メロディも評価してしまう点

スキヤット学習部では本来、メロディに適した言葉だけの評価で、応答メロディ生成部によって生成されたメロディ自体は考慮しない。しかし、心理的にメロディ自体が悪いとそれにつられて悪い評価をしてしまいがちになる。提案システムではメロディと言葉を完全に分けて学習させたが、メロディも報酬によって強化学習できるようになれば、この問題を解決できると考えられる。

### 明示的に状態を決めている点

状態は、音階と音の長さを明示的に離散化して決めてしまっていたので、Q 関数をニューラルネットなどを用いて関数近似することでユーザの好みをより柔軟に学習しやすくなると考えられる。

## 初音ミクが意識しにくい点

提案システムでは、初音ミクのぬいぐるみからスキヤット音が流れるようにしたが、動かない人形と歌声だけではセッション相手である初音ミクを意識しにくかった。ディスプレイ上に初音ミクを表示し、スキヤット時にリップシンクさせるといったような動きを加えることで初音ミクがスキヤットを歌っているという感覚がより強まると考えられる。

## 6. まとめ

本研究では、ユーザ好みのスキヤットを強化学習によって学習していくジャムセッションシステムを提案し、実装を行った。提案システムは応答メロディ生成部とスキヤット生成部に大きく分けられ、応答メロディ生成部では、先行研究の「A.I. Duet」を用いた。スキヤット生成部では、Q 学習を用いてユーザの好みのスキヤットを学習させた。最後に、著者が提案システムを動作確認することで、ユーザ好みのスキヤットが学習できていることを確認した。しかし、それと同時に課題点も多く見つかると、それに対する改善策を検討した。今回は著者 1 人での実験であったため十分な情報が得られなかった。今後は、今回見つけた課題点の改善と共に、複数の被験者を用いた評価実験によって、より詳しく検証を行う必要があると考えられる。

## 参考文献

- [1] Nishijima, M. and Watanabe, K.: Interactive Music Composer based on Neural Networks, *ICMC*, pp. 53–56 (1992).
- [2] 後藤真孝, 日高松本 英明, 黒田洋介, 村岡洋一: 仮想ジャズセッションシステム: VirJa Session, *情報処理学会論文誌*, Vol. 40, pp. 1910–1921 (1999).
- [3] 青野裕司, 片寄晴弘, 井口征士: バンドライクな音楽アシスタントシステムについて, *情報処理学会研究報告音楽情報科学 (MUS)*, Vol. 1994, pp. 45–50 (1994).
- [4] 北原鉄朗, 戸谷直之, 徳網亮輔, 片寄晴弘: BayesianBand: ユーザとシステムが相互に予測し合うジャムセッションシステム, *情報処理学会論文誌*, Vol. 50, pp. 2949–2953 (2009).
- [5] 西嶋正子, 村上公一: ニューロ・ドラマー, 計測と制御, Vol. 30, pp. 344–347 (1991).
- [6] Yotam Mann: A.I. Duet, (オンライン), 入手先 < <http://experiments.withgoogle.com/ai/ai-duet> > (参照 2018-2-5)