The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

ユーザ好みのスキャ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の開発

Development of a Jam Session System with Hatsune Miku That Learns User's Favorite Scat by Reinforcement Learning

鶴田 穣士 Joji TSURUDA　岡 夏樹 Natsuki OKA　田中 一晶 Kazuaki YANAKA

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

Musicians often use jam sessions as a way to communicate with music. Many studies on jam session systems, which enable jam sessions between people and machines, have been made. In jam sessions, singers often use the scat singing. In our proposed system, a user plays keyboard and the system will reply in scat singing of Hatsune Miku according to the user's performance. We aim to make the system learn the user's preferred scat words by reinforcement learning through jam sessions. We implemented the learning system and confirmed that the system could learn the user ' s preference in scat words.

1. はじめに

近年，家庭用コミュニケーションロボットなどの普及に伴い，人間と機械とのコミュニケーションがより一般化しつつある．音楽の分野においても，人間の代替として機械がコミュニケーション相手として使われるようになってきている．機械による自動演奏を人間が聴くという形態としては，初音ミク *1 などのバーチャル・アイドル主演のライブコンサートが催されている．透明なスクリーン上に映された 2 次元のアイドル初音ミクに対して観客たちは，人間相手と同じように手を振り声援を送る．それに対して初音ミクは手を振り返したり「ありがとう」と言い返したりする．

しかし，これらは一見，機械が声援に対する応答をしている，つまり人間と機械のインタラクションができているように見えて，実際は，観客の反応に関係なく事前にプログラミングされた通りのことをしているにすぎない．

よりインタラクティブな音楽を用いたコミュニケーションとして，ジャムセッションと呼ばれるものがある．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演奏者間のインタラクションが顕著である点が他の音楽を用いたコミュニケーション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．

また，ジャムセッションでは，ボーカルが声を楽器のように使う歌唱法であるスキャットが使われる場合がある．スキャットとは，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ある．本研究では，近年の歌声合成技術の発達を受け，スキャットを用いた初音ミクとのジャムセッションシステムを提案する．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，ただ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，コミュニケー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．

連絡先: 京都工芸繊維大学 大学院工芸科学研究科 情報工学専攻 インタラクティブ知能研究室，tsuruda@ii.is.kit.ac.jp

*1 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．

2. 目的

本研究では，ジャムセッションに馴染みのない演奏者でも気軽に始められるような初音ミクとのジャムセッションシステムの開発を目指す．対コンピュータであれば，初心者でも人間相手より，臆せず気軽に始められ，その面白さに気付き，興味を持つきっかけになり得ると考えられるからである．スキャ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分かれやすい．よって，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ャ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．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，人間のスキャットの型にとらわれない，自由で新しいスタイルのスキャットの生成も期待できると考えられる．

3. 関連研究

ジャムセッションシステムは，これまで様々な研究が行われてきた．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 [Nishijima 92] [後藤 99] や，ユーザに合わせてコード進行を変えるもの [青野 94]，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 [北原 09] 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 [西嶋 91]．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 [Yotam 17]．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

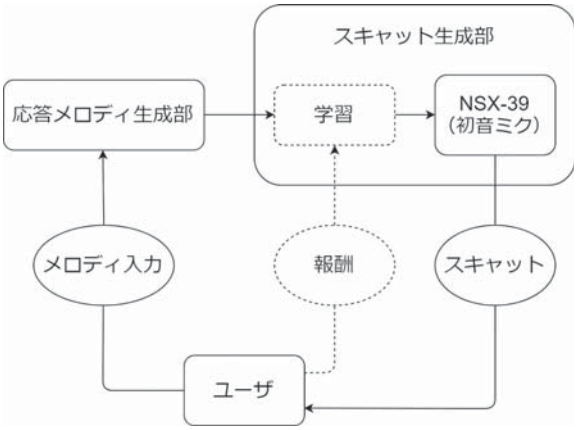

図 1: 提案システムの概要図

図 2: 提案システムを使用している様子

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．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．

4. 提案システム概要

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キャ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，スキャットのメロディを生成する．
3. スキャット生成部にて，生成されたメロディと学習経験をもとに初音ミクがスキャットとして歌う．
4. ユーザはそのスキャ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キャットを学習し，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キャット音が流れる．現時点までの開発は，主にスキャ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 5. 節で簡単に紹介し，スキャット生成部は 6. 節にて詳しく説明する．

5. 応答メロディ生成部

先行研究の「A.I. Duet」を利用して，メロディ生成やディスプレイ表示を行う．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，スキャット生成部に送る．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．

6. スキャット生成部

6.1 スキャット生成部の概要

スキャットには，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．スキャット生成には，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．スキャ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，初音ミクがスキャットを歌う．次に，ユーザはそのスキャットを自分の好みかどうかで評価し，図 2 中下部のように PC キーボード入力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キャットを強化学習していく．

6.2 学習方法

提案システムでは，強化学習手法の一つである Q 学習を用いた．

（1）行動

スキャ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．

- 一般的にスキャットに使われる言葉と使われない言葉を含む．

*2 http://magenta.tensorflow.org/

- 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む.
- 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）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．そのためスキャット 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．スキャ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.

（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.

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

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.

（4）報酬

ユーザは初音ミクのスキャット（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．良いと判断した場合は +1，悪いと判断した場合は −1 の報酬を与え，良くも悪くもないと判断した場合はユーザは何もする必要は無く，その場合報酬は与えられない.

（5）Q 値更新

報酬が与えられたときに，対象となるスキャ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.

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r - Q(s,a)) \tag{2}$$

ただし，r は報酬であり，学習率 $\alpha = 0.1$ とした．また，Q 値の初期値は全て 0 とした.

7. 実験

7.1 実験環境

執筆者の 1 人によって学習動作確認実験を行った．メロディ入力からスキャット出力までを 1 エピソードとして，250 エピソード行った.

表 1: 学習後の Q 値表

状態 音の高さ	状態 音の長さ	状態 前の行動 \ 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	***-0.570***
		「ビ」	0.122	0.046	-0.386	-0.293
		「ケ」	***-0.506***	-0.263	-0.344	***-0.522***
		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」	***-0.613***	***-0.453***	-0.410	***-0.556***
		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	***-0.613***	***-0.522***
		無し	0.399	0.501	***-0.445***	***-0.530***

表 2: 言葉の遷移に着目した学習後の Q 値の平均値

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」	***-0.435***	-0.210	***-0.389***	***-0.486***
無し	0.180	0.128	-0.284	-0.382

7.2 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ャットが出力されるように感じられた.

7.3 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」が続くスキャットとなっている．これは一般的によく使われるスキャット「ドゥビ」と言葉は同じである．しかし，一般的なスキャットでは「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のスキャット生成ができたのではないかと考えられる．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

キャ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だったといえる．

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ャットが好みであったと考えられる．

8. 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

スキャ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．

状態を粗く離散化した点

人とのインタラクションを通した学習を可能とするため，状態数を抑える必要があったため，状態は各時点の音の高さと音の長さを極めて粗く離散化したものとした．このため，音の高さや長さに応じた言葉の学習ができたとは言い難い結果となった．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離散化を不要にして，音の高さや長さに応じたユーザの好みを学習できる可能性がある．さらに，音の高さや長さの変化も考慮した好みの学習を可能とすることも検討したい．

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ャ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，スキャ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ャ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．

9. まとめ

本研究では，ユーザ好みのスキャ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ャ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」 [Yotam 17] を用いた．スキャット生成部では，Q 学習を用いてユーザの好みのスキャットを学習させた．執筆者の 1 人が提案システムを試用することで，ユーザ好みのスキャットが学習できることを確認した．しかし，それと同時に課題も多く見つかり，それに対する改善策を検討した．今回は執筆者の 1 人だけによる実験であったため十分な評価情報が得られなかった．今後は，今回見つかった課題点の改善と共に，複数の一般被験者を用いた評価実験を実施する必要がある．

参考文献

[Nishijima 92] Nishijima, M. and Watanabe, K.: Interective Music Composer based on Neural Networks, *ICMC*, pp. 53–56 (1992)

[Yotam 17] Yotam Mann : A.I. Duet, (2017),〈 http://experiments.withgoogle.com/ai/ai-duet〉(参照 2018-2-5)

[後藤 99] 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)

[西嶋 91] 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)

[青野 94] 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)

[北原 09] 北原 鉄朗, 戸谷 直之, 徳網亮輔, 片寄 晴弘:Bayesian-Band : 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)