Vol.2021-MUS-131 No.29
Vol.2021-SLP-137 No.29
2021/6/18

楽曲の音響信号予測学習に基づく楽曲への選好形成のモデル化

吉永 塁[1,a)]　岡 夏樹[1]　田中 一晶[1]

概要：エージェントが音楽選好を持つことで，人間とエージェントの共同音楽聴取におけるインタラクションはより深く充実したものになると考えられる．本研究では人間と共同音楽聴取を行うエージェントのための楽曲への選好形成モデルを提案する．音楽への好感度は親近性によるという知見に基づき，エージェントの音楽選好を「楽曲の続きの予測しやすさの学習に伴う変化」により表現する．提案システムを試作し，その可能性について検討を進めており，その途中成果を報告する．

1. 緒言

友人や家族の一員として扱われ，使い続けられるようなロボットやエージェントは未だ実現していない．エージェントが人間と共生関係を築くためには，人間の命令を受動的に遂行するだけでなく，主体性を持ち自己主張できる能力が必要であると考える．ここで，スマートスピーカの普及を踏まえ，人間とエージェントの共同音楽聴取の場面を想定し，エージェントの音楽選好に着目する．エージェントが音楽選好を持つことで，例えば，エージェントが好きな音楽を人間に薦めたり，人間との共同音楽聴取において聴取した音楽の感想を伝えたり，あるいは選曲時に自身の選好をもとに聴きたい曲を主張するといった働きかけが可能となる．さらに，音楽聴取の経験により選好を形成する能力をエージェントに持たせることで，エージェントの音楽選好が，共同音楽聴取の相手の選好へと近づいていく過程を表現できると考える．これにより，長く連れ添った夫婦が経験を共有することにより似たような選好を持つようになるといった傾向を表現できるのではないかと考える．以上より，本研究では，人間と共同音楽聴取を行うエージェントのための楽曲への選好形成モデルを提案する．

2. 関連研究

2.1 音楽選好

Madison ら [1] は，音楽への好感度（Liking）に影響する要因を調査している．リスニング実験により，音楽への好感度は，音楽の複雑さ（Complexity）には依らず，繰り返し聴取することにより単調に増加することを発見している．また，結論として，似た音楽を聴いたことがあるかという意味での音楽への親近性（Familiarity）こそが，好感度を決定する唯一の最重要変数であることを導いている．

2.2 音楽生成

音楽生成の研究は多数行われているが，近年，音楽を音響信号として生成する手法が大きな発展を見せている．特に，Dhariwal らが提案した Jukebox[2] と呼ばれるモデルは，数分に及ぶ長さでの多様な楽曲の生成を可能としている．以下，Jukebox に用いられているアーキテクチャについて説明する．

Jukebox は Music VQ-VAE と Music Prior により構成される．Music VQ-VAE は，VQ-VAE[3] を階層化した Hierarchical VQ-VAE[4] を音楽へと適用したものである．3 段階の階層構造を持ち，各階層の VQ-VAE はそれぞれ異なる圧縮率（上位の階層ほど高い圧縮率を持つ）で音響信号とその離散圧縮表現の対応関係を学習する．また，この音響信号の離散表現はコード（code）と呼ばれる．Music Prior は，VQ-VAE と対応する 3 階層の構造を持ち，各階層の VQ-VAE が作る離散空間上で自己回帰的にコードを生成する．コードの生成過程において，最上位の Prior が基本となるコードを生成し，下位の Piror によりアップサンプリングが行われる．Prior により生成されたコードを VQ-VAE で変換することにより，音響信号が得られる．

[1] 京都工芸繊維大学
Kyoto Institute of Technology, Matsugasaki hashigami-cho 1, Sakyo-ku, Kyoto 606-8585, Japan
[a)] yoshinaga@ii.is.kit.ac.jp

Vol.2021-MUS-131 No.29
Vol.2021-SLP-137 No.29
2021/6/18

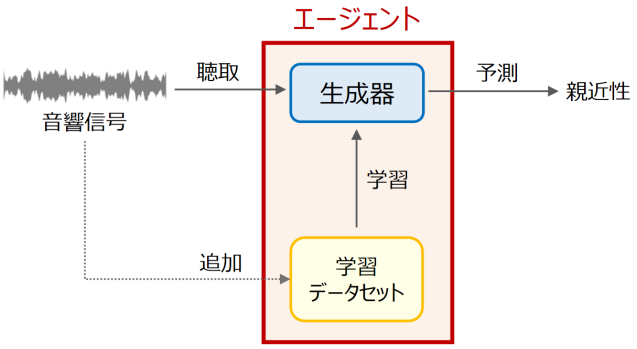

図 1 システム概要図：**予測**：楽曲を音響信号として入力し，その続きを生成する．生成した続きと実際の続きを比較することにより予測誤差を計算し，楽曲への親近性として出力する．**学習**：過去に入力された音響信号の集合をもとに，楽曲の続きの生成を担う生成器の学習を行う．

3. 提案手法

3.1 親近性の定量的評価

本研究の目的は音楽選好の形成過程のモデル化である．2.1 節に示した先行研究の知見に従い，音楽への選好は親近性によるとする．よって本研究では，楽曲への親近性を判別できるエージェントを構築する．ここで楽曲への親近性の定量的な評価方法について検討する．

楽曲への親近性を定量的に評価する方法について，単純なものとしては，楽曲ごとにその聴取回数を数え，その数の大小を親近性として評価するという方法が考えられる．しかし，楽曲への親近性の判別には，その楽曲を聴いたことがあるかの判別に加え，それに似た楽曲を聴いたことがあるかの判別も必要となる．よって，単純に聴取回数を数えるだけでは，エージェントが親近性のある楽曲を判別できず，不適であると判断した．また，他の方法としては，聴いた楽曲を埋め込み表現へと変換し，埋め込み空間において過去に聴いた楽曲の集合と今聴いている楽曲との距離を親近性としてみなすという方法が考えられる．しかし，過去に聴いた楽曲の集合は大きくなり続け，さらに楽曲の長さはそれぞれ異なる長さを持つ．よって，距離の計算において計算量が爆発することが予想されるため，不適であると判断した．以上を勘案し，本研究では予測による親近性の評価を提案する．この手法について，以降で説明する．

3.2 システム

システムの概要図を図 1 に示す．エージェントが聴取した楽曲への親近性を判別する流れは以下の通り．

(1) 入力：楽曲を音響信号として入力する．
(2) 予測：過去に聴いた曲をもとに，入力された楽曲の続きを生成する．生成した続きを実際の続きと比較し予測誤差を計算する．
(3) 出力：予測誤差の値を楽曲への親近性として出力する．

本システムの特徴として，入力で必要とするのが音響信号（WAV 形式の楽曲ファイル）のみであるという点が挙げられる．これにより，例えば，MIDI 形式が利用可能な楽曲のみを対象とするなどの制限を課すことなく，より広範囲の音楽を対象とすることが可能である．本システムでは，2.2 節で述べた Jukebox[2] の構造を利用する．

3.2.1 システム：予測

本システムにおいて，予測は次の 3 部で構成される（図 2）．

(1) 変換：VQ-VAE により，入力された音響信号を離散コードへと変換する．このとき，VQ-VAE は学習済みのものを使用する．
(2) 生成：Prior により，(1) で得られたコードの続きを生成する．
(3) 比較：生成した続きと実際の続きの比較し，予測誤差を計算する．比較の指標には交差エントロピーを用いる．

本システムにおける楽曲の続きの生成では，計算量削減のためアップサンプリングの過程は省略した．これにより，下位の階層が持つ局所的な情報が失われることが予想されるが，計算量の削減を優先し，Jukebox の持つ構造のうち，最上位の階層のみを用いた．また，以降の説明において，Jukebox の Prior に相当する部分を単に生成器と呼ぶ．

3.2.2 システム：学習

本システムでは，生成器の学習データとして，エージェントが過去に聴いた楽曲の集合を与える．生成器の学習は，学習データ（楽曲）の尤もな続きの生成を目的として行われる．よって，エージェントが過去に聴いたことのある楽曲やそれに似ている楽曲は予測されやすく，聴いたことがなく馴染のない楽曲は予測されにくくなる．これにより，エージェントの楽曲への親近性が，その楽曲の続きの予測のしやすさにより表現できると考えた．

一般に，学習データに過剰適合するとモデルの汎化性が失われるため過学習は避けるべきである．しかし，楽曲への親近性の判別が本研究の目的であるため，エージェントが聴いたことのない楽曲（未知のデータ）に対する予測精度の向上を目指すのではなく，聴いたことのない楽曲の予測精度はむしろ下げて，聴いた曲との差を大きくしたい．そこで，意図的に過学習を起こすべきだと考えた．このため，生成器の学習において，過学習を防ぐ効果を持つ重み減衰（Weight Decay）が無効になるように設定する．

4. 実験

4.1 仮説

本研究で構築するエージェントは楽曲への親近性を判別するように設計されたものである．次のように仮説を立て，実験により検証する．

Vol.2021-MUS-131 No.29
Vol.2021-SLP-137 No.29
2021/6/18

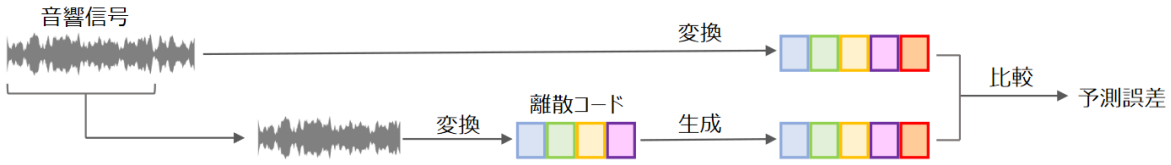

図 2 予測の過程：(1)Music VQ-VAE により入力された音響信号を離散コードへと変換し，(2) 生成器によりそれに継続する新たなコードを生成する．(3) 生成したコードを実際のものと比較し，予測誤差を計算する．

仮説：予測誤差は，(1) 聴いたことのある楽曲，(2) それに似た楽曲，(3) 聴いたことがなく馴染のない楽曲，の順で大きくなる．

4.2 実験

4.1 節で示した仮説の検証のため，エージェントを実装し，以下の手順で実験を実施した．

手順 1： 楽曲を 1 曲選択し，その楽曲を用いて生成器の学習を行う．事前に学習曲線を確認し，充分に学習が進んだと考えられる 4000 エポックまで学習を行うこととした．

手順 2： 生成器の学習後，学習曲を含む全楽曲に対し予測を行う．各楽曲の冒頭から終わりまで 1 秒ごとにその続きの約 24 秒を予測する．

実験で使用した楽曲を表 1 に示す．以降の説明において，各楽曲はラベル名により示し，特に学習に用いられた楽曲を学習曲と表記する．

学習曲の特徴に起因する差異の確認のため，以下の 2 条件について実験を実施した．

(i) クラシックの楽曲による学習 学習曲として，Classic-1 を選択する．Classic-1, Classic-2 が同一の作曲家による楽曲であることより，予測誤差の順序は，Classic-1 < Classic-2 < その他の楽曲 となると予想した．

(ii) ロックの楽曲による学習 学習曲として，Rock-1 を選択する．(i) と同様にして予測誤差の順序は，Rock-1 < Rock-2 < その他の楽曲 となると予想した．

5. 結果と考察

5.1 結果

実験結果を図 3, 図 4 に示す． Classic-1 を学習曲とした場合では，Classic-3 を除くと概ね仮説に沿う結果が得られた．また，Rock-1 を学習曲とした場合では，Classic-3 を除くと学習曲は最小の値を取ったが，学習曲と同一のアーティストによる楽曲である Rock-2 がクラシックの楽曲よりも大きな値を取る結果となった．

5.2 考察

実験において特徴的な動きを示した Classic-3 に着目する．この楽曲の予測誤差はその他の楽曲の予測誤差に比

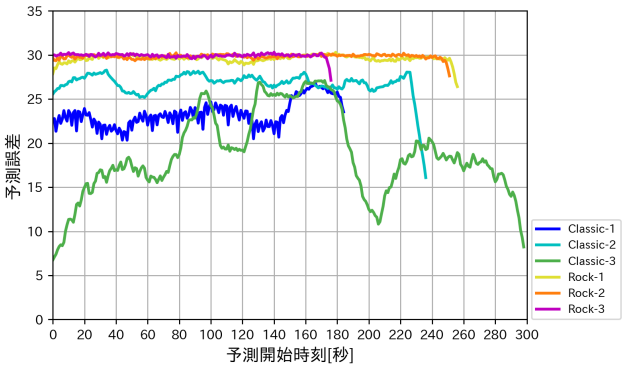

図 3 実験結果（学習曲：Classic-1）：Classic-3 を除くと概ね仮説通りとなった．

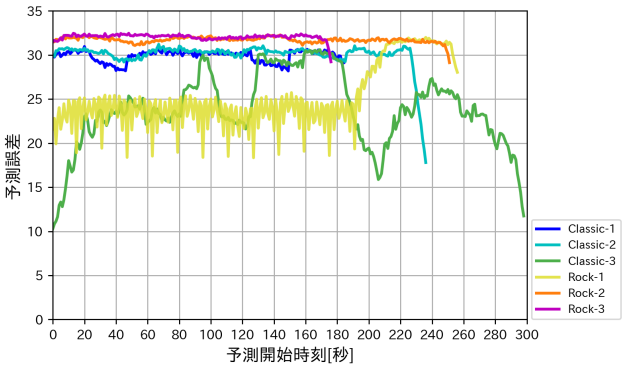

図 4 実験結果（学習曲：Rock-1）：Classic-3 を除くと，学習曲である Rock-1 の値は最小となったが，学習曲と同一のアーティストの楽曲である Rock-2 がクラシックの楽曲よりも大きな値となった．

べ，広範囲の値を取る．楽曲の序盤・終盤では学習曲より小さい値を示し，中盤では学習曲以外の楽曲と同程度の値を取る．ここで，この楽曲の曲調・雰囲気に着目すると，Classic-3 は序盤・終盤において無音や持続音が多く，中盤では音の数が増え，動きのある構成となっている．これより，楽曲内において，無音やそれに近い持続音・減衰音ほど予測が容易であることが推察できる．さらに学習曲以外の楽曲の予測誤差に着目すると，両方の条件において，クラシックの楽曲とロックの楽曲で値が分離していることが読み取れる．Classic-3 についても，中盤ではその他のクラシックの楽曲と同程度の値を取る．クラシックの楽曲はピアノにより，またロックの楽曲はバンドにより演奏されて

表 1　実験で使用する楽曲一覧

ラベル	曲名	アーティスト（作曲家）名	ジャンル	楽器編成	秒数
Classic-1	Partita 1-2.Allmande BWV.825	J.S.Bach	Classic	ピアノ	209
Classic-2	Partita 5-7.Gigue BWV.829	J.S.Bach	Classic	ピアノ	261
Classic-3	月の光	C.A.Debussay	Classic	ピアノ	323
Rock-1	KICK IT OUT	BOOM BOOM SATELLITES	Rock	バンド	281
Rock-2	9 Doors Empire	BOOM BOOM SATELLITES	Rock	バンド	276
Rock-3	透明少女	NUMBER GIRL	Rock	バンド	201

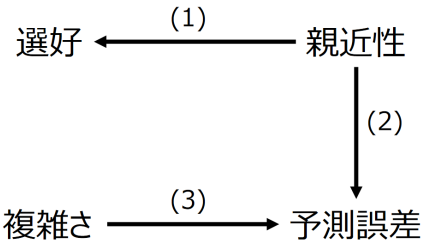

図 5　選好・親近性・予測誤差・複雑さの関係性：(1) 先行研究より，楽曲の選好はその楽曲への親近性によることが得られている．(2) 学習曲の予測誤差は他の楽曲よりも小さくなることより，楽曲への親近性（学習量）は予測誤差へ影響する．(3) 本実験より，楽曲の複雑さが予測誤差へ影響することが推察される．

いることを踏まえると，楽曲間においても，楽器編成が単純なもの，つまり楽曲を構成する音がより無音に近いものほど予測が容易になることが推察される．無音に近い音ほど予測が容易になることの理由として，無音や持続音・減衰音は楽曲に関わらず現れる，音楽に一般的な概念であるためであると考えられる．

前述の「無音に近いほど予測しやすい」という仮説が正しいとすれば，楽曲の複雑さが予測精度へ影響する．また，学習曲の予測誤差はその他の楽曲の予測誤差よりも小さい値を取ることから，予測誤差は楽曲への親近性（学習量）にも影響を受ける（図 5）．よって，予測誤差から親近性を判別するには，さらに楽曲の複雑さも考慮に入れる必要がある．

6. 課題

考察で述べた通り，予測誤差の値が楽曲の持つ複雑さに影響を受けることが伺える．よって，親近性を判別するには，予測誤差に加え楽曲の複雑さを考慮する必要があり，楽曲の複雑さを定める評価指標が必要となる．さらに，予測誤差と楽曲の複雑さの両方から親近性を判別する手法について検討が必要である．

本実験では，エポック数を固定し，選好形成の過程における一時点の選好にのみ注目したが，楽曲を聴取することにより選好が変化していく様子を表現する必要がある．本手法では予測誤差により選好を表現するため，予測誤差の変化と好感度の変化をどのように対応付けるかについて検討が必要である．

7. 結言

本研究ではエージェントの音楽選好に着目し，楽曲への選好形成のモデル化について検討した．先行研究による知見に従い，音楽選好は親近性によると定め，楽曲の続きの予測により楽曲への親近性を判別する手法を提案した．提案システムを試作し動作検証実験を行った．親近性があると判断されるべきである学習曲と同一の作曲家・アーティストによる楽曲の判別が難しい点や，楽曲の複雑さが予測に影響を与えてしまう点など検討の余地は残るが，学習曲の判別に関してはある程度達成されたといえる．前述の課題点について検討・修正した上で，選好形成の過程の表現手法を検討・構築することが今後の課題である．

謝辞　本研究は JSPS 科研費 JP20H05564 の助成を受けたものである．

参考文献

[1] Madison, G. and Schiölde, G.: Repeated Listening Increases the Liking for Music Regardless of Its Complexity: Implications for the Appreciation and Aesthetics of Music, *Frontiers in Neuroscience*, Vol.11, p.147 (online), DOI: 10.3389/fnins.2017.00147 (2017).

[2] Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radfordd, A., and Sutskever, I.: Jukebox: A Generative Model for Music, *arXiv preprint arXiv:2005.00341* (2020).

[3] van den Oord, A., Vinyals, O. and kavukcuoglu, k.: Neural Discrete Representation Learning, *Advances in Neural Information Processing Systems 30* (Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R., eds.), Curran Associates, Inc., pp.6306–6315 (online), available from ⟨http://papers.nips.cc/paper/7210-neural-discrete-representation-learning.pdf⟩ (2017).

[4] Razavi, A., van den Oord, A. and Vinyals, O.: Generating Diverse High-Fidelity Images with VQ-VAE-2, *Advances in Neural Information Processing Systems 32* (Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché–Buc, F., Fox, E. and Garnett, R., eds.), Curran Associates, Inc., pp. 14866–14876 (online), available from ⟨http://papers.nips.cc/paper/9625-generating-diverse-high-fidelity-images-with-vq-vae-2.pdf⟩ (2019).