7T-04

ゲームシーンからそれに適したBGMの音響特徴を予測する手法の検討

林龍星 † 北原鉄朗 †
† 日本大学文理学部情報科学科

1. はじめに

ビデオゲームにおいて，BGMは雰囲気を演出する上で重要な役割を担う．そのため，ゲームやシーンの雰囲気とBGMの間には強い関係があると考えられる．

そこで本研究では，ユーザが自作ゲームにふさわしいBGMを探索することを想定し，シーン（ゲームの場面を表す情景）および参考にしたいゲームを指定するとそれにふさわしいBGMを探索するシステムを提案する．

ゲームのための音楽生成に関する研究は，RNNを用いて乱数から楽曲を生成するもの[1]，ファミコンゲーム音楽のデータベースに関する研究[2]などがあるが，シーンを入力して楽曲を生成するものではない．映像やダンスシーンから音楽を生成するもの[3][4]もあるが，ゲームを対象としていない．

本システムを実現する上での課題は，シーンの入力方法である．シーンは自然の様子，建物の様子，人物，人物の行動などの様々な要素で構成されるが，BGMとシーンとがペアになったデータを探すのは容易ではない．

本稿では，シーンとしてゲーム内の映像そのものを入力として用いる．学習データにはインターネット上の動画共有サイトに大量に投稿されているゲームのプレイ動画を用いる．

2. 提案手法

本システムは，BGMを付加したいゲームの映像が与えられ，機械学習を用いて，そのシーンに適したBGMの音響特徴を予測する．機械学習モデルはゲームごとに学習してあるため，BGMを模倣したいゲームを選ぶことができる．その後，BGM候補として用意されたフリー音源集から，予測された音響特徴に最も近いものを探索する．

2.1 入力・出力データの前処理

入力・出力データには，YouTube上に投稿されているspeedrun動画を用いる．この動画をMP4形式で保存したものを12秒ごとに分割し，次の処理を行う．この処理で得られるデータをゲームごとに100個用意し，シャッフルしてから半分を学習データ，残りをテストデータにする．

2.1.1 入力データ

12秒ごとに分割されたMP4ファイルに対して，OpenCVのcvtColor関数で画象をRGBに変換する．さらに，OpenCVのresize関数で画像のサイズを80:80に変更する．これにより，80:80:3のデータが360フレーム得られる．

2.1.2 出力データ

入力データと同じMP4ファイルからオーディオ部をWAV形式で抜き出し，librosaを用いてフレームごとにstft, cqt, iirt, salience, chroma_stft, chroma_cqt, chroma_cens, mel-spectrogram, mfcc, delta, nmfを抽出する．

2.2 CNN-LSTMによるBGMの音響特徴の予測

上で述べた入力データから出力データを予測するモデルをCNN-LSTMにより構築する（図1）．なお，最適化手法にはADAMを用いる．損失関数には平均二乗誤差を用いる．バッチサイズは16とする．chroma系の音響特徴を用いる場合のエポック数は500，それ以外の音響特徴を用いる場合のエポック数は100とする．

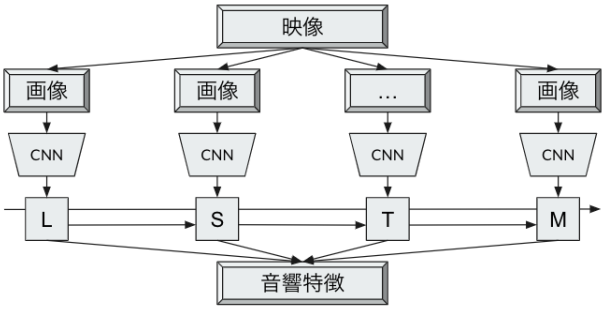

図1 CNN-LSTMの構成

表1 評価実験に使用するゲーム

学習用	Undertale, クロノ・トリガー
入力用	Ghost of Tsushima, OFF, OMORI, すばらしきこのせかい, ニーア オートマタ, ペルソナ5, モンスターハンター ストーリーズ, ロマンシング サ・ガ3, ワイルドアームズ, 大神

2.3 フリー音源の出力

予測された音響特徴に最も近いBGMを，事前に作成したフリー音源集から探索する．フリー音源集の各楽曲から全音響特徴を抽出し，予測した音響特徴と各楽曲との距離をEarth Mover's Distanceにより計算し，昇順に結果を出力する．

3. 評価実験

提案手法について評価実験を行う．提案手法はユーザが自作ゲームに類似したゲームを学習したモデルを選択することを想定しているため，この状況を疑似的に再現するために学習するゲームに類似した既存のゲームを求める．次に，予測された音響特徴の妥当性を，予測された音響特徴と実際のBGMの音響特徴を比べることで検証する．最後に出力されたBGMの妥当性を定性的に論ずる．

3.1 データセット

評価実験に用いるゲームはYouTube上に投稿されているspeedrun動画を用いる（表1）．各ゲームから，12秒分のシーンを6個抽出する．内訳は戦闘シーンが2個，探索シーンが2個，会話シーンが2個である．BGM探索用のフリー音源はフリー音源サイト『bensound』，『DOVA-SYNDROME』，『魔王魂』から得た50曲のWAVファイルを用いる．

3.2 ゲームの類似度に関する実験

学習に使用したゲーム2作品に対して類似度が高いゲームを求め，以後の評価データとして用いる．ゲームの類似度はシーン（画像）の距離を画像のハッシュ値の差分として求め，その平均値を取ることで算出する．その結果,『Undertale』に最も類似しているゲームは『OMORI』,『クロノ・トリガー』に最も類似しているゲームは『ロマンシング サ・ガ3』であ

Method for predicting the acoustic characteristics of suitable background music from game scenes
by Ryusei Hayashi and Tetsuro Kitahara (Nihon University)

表 2 『OMORI』に対する音響特徴評価

実際 \ 予測	戦闘シーン	探索シーン	会話シーン
戦闘シーン	0.4425	0.3215	0.6135
探索シーン	0.3104	0.1513	0.2861
会話シーン	0.4058	0.1603	0.2561

表 3 『ロマンシング サ・ガ 3』に対する音響特徴評価

実際 \ 予測	戦闘シーン	探索シーン	会話シーン
戦闘シーン	0.7504	0.3243	0.5270
探索シーン	0.7279	0.2997	0.4128
会話シーン	0.8159	0.4510	0.4849

ることが示された.

3.3 予測された音響特徴の妥当性に関する実験

3.3.1 方　　法

『Undertale』と『クロノ・トリガー』を対象に，11 種類の音響特徴を別々に予測するモデルを学習し，前者には『OMORI』，後者には『ロマンシング サ・ガ 3』から抽出したシーンを入力して各音響特徴を予測する．シーン s に対する i 種類目の音響特徴の予測値を $y_i^{\mathrm{pred}}(s)$，同じシーンの実際の BGM の音響特徴を $y_i^{\mathrm{true}}(s)$ とする．このとき，これらの距離 $\mathrm{dist}(y_i^{\mathrm{pred}}(s), y_i^{\mathrm{true}}(s))$ は，他のシーンとの距離 $\mathrm{dist}(y_i^{\mathrm{pred}}(s), y_i^{\mathrm{true}}(s'))$ $(s \neq s')$ より小さい方が望ましい．そこでこれらを比較する．この値は用いる特徴によって変化するため，11 種類の音響特徴の組合せ（2048 個）に対して，

$$\sum_{s\in S}\sum_{s'\in S}\{\mathrm{dist}(y_i^{\mathrm{pred}}(s'), y_i^{\mathrm{true}}(s)) - \mathrm{dist}(y_i^{\mathrm{pred}}(s), y_i^{\mathrm{true}}(s)\}$$

（S: シーンの集合）が最大になる組合せ i を求め，それを用いる．距離には EMD を用いる．

3.3.2 結果・考察

『OMORI』に対する結果と『ロマンシング サ・ガ 3』に対する結果をそれぞれ表 2, 3 に示す．前者は cqt と chroma_cens の組合せ，後者は salience と delta の組み合わせである．『OMORI』に対しては，探索シーン同士，会話シーン同士の音響特徴は他のシーンよりも距離が小さくなったが，戦闘シーン同士はそのような結果にはならなかった.『ロマンシング サ・ガ 3』に対しては探索シーン同士の音響特徴は他のシーンよりも距離が小さくなったが，戦闘シーン同士，会話シーン同士はそうならなかった.

3.4 出力された BGM の妥当性

表 4 より,『Take a Chance !』は電子音の音色で緊張した雰囲気を演出しているため，戦闘シーンに適していると考えられる.『Shall we meet?』は繰り返しの多いメロディーで穏やか雰囲気を演出しているため，戦闘シーンには適していないと考えられる.『Dew』は速いテンポで軽快な雰囲気を演出しているため，探索シーンに適していると考えられる.『Downtown』はジャズの曲調で落ち着いた雰囲気を演出しているため，会話シーンに適していると考えられる.『オーケストラ 24』はコーラスで緊張した雰囲気を演出しているため，一部の会話シーンに適していると考えられる.

表 5 より,『サイバー 42』は特徴的な楽器で軽快な雰囲気を演出しているため，一部の戦闘シーンには適していないと考えられる.『Badass』は変化の少ない曲調で落ち着いた雰囲気を演出しているため，戦闘シーンには適していないと考えられる.『いけないドーナッツ』は音符の少ないメロディであやしい雰囲気を演出しているため，一部の探索シーンに適していると考えられる.『Badass』は変化の少ない曲調で落ち着いた雰囲気を演出しているため，探索シーンには適していないと考えられる.『8bit28』は変化の大きい曲調でせわしない雰囲気を演出しているため，一部の会話シーンに適していると考えられる.

表 4 『クロノ・トリガー』を chroma_stft で学習して『ロマンシング サ・ガ 3』を入力したときに出力された BGM

BGM \ シーン	戦闘	戦闘	探索	探索	会話	会話
Shall we meet?	0.08	0.16	0.21	0.16	0.13	0.36
Take a Chance !	0.19	0.06	0.17	0.21	0.04	0.38
Dew	0.21	0.42	0.11	0.05	0.33	0.14
オーケストラ 24	0.21	0.17	0.16	0.17	0.03	0.31
Downtown	0.33	0.38	0.16	0.10	0.29	0.13

表 5 『Undertale』を chroma_cens で学習して『OMORI』を入力したときに出力された BGM

BGM \ シーン	戦闘	戦闘	探索	探索	会話	会話
サイバー 42	0.14	0.26	0.26	0.15	0.32	0.33
Badass	0.24	0.11	0.37	0.33	0.16	0.25
いけないドーナッツ	0.31	0.37	0.23	0.16	0.49	0.33
8bit28	0.29	0.44	0.24	0.33	0.15	0.19

『Undertale』を chroma_cens で学習したモデルはほぼ全てのシーンにおいて『禁域にて』,『魔女の小部屋』,『ファンタジー 06』のどれかを最も適切なフリー音源に選んだ.『禁域にて』は不協和音で不気味な雰囲気を演出しているため,『OFF』に適していると考えられる.『魔女の小部屋』は西洋楽器であやしい雰囲気を演出しているため，探索シーンに適していると考えられる.『ファンタジー 06』はコーラスで壮大な雰囲気を演出しているため,『ワイルドアームズ』に適していると考えられる.

4. おわりに

本稿では，ゲームシーンからそれに適した BGM の音響特徴を予測する手法を検討した．この手法では，BGM を模倣したいゲームを機械学習したモデルに BGM を付与するシーンを与え，そのシーンに適した BGM の音響特徴を予測する．その後，予測した音響特徴に最も類似した音響特徴を持つフリー音源を探索する．この手法について予測された音響特徴の妥当性と出力された BGM の妥当性を評価した．今後は，新規ゲームを開発する環境において実働試験を行った上でシステムを改良する必要がある．

謝辞 本研究は，科研費 22H03711, 21H03572 の支援を受けた.

参 考 文 献

1) Nicolas Mauthes：“VGM-RNN: Recurrent Neural Networks for Video Game Music Generation”, SJSU, 2018.
2) Chris Donahue, Huanru Henry Mao, Julian McAuley：“The NES Music Database: A Multi-Instrumental Dataset with Expressive Performance Attributes”, ISMIR, pp.475–482, 2018.
3) Chuang Gan, Deng Huang, Peihao Chen, Joshua B. Tenenbaum, Antonio Torralba：“Foley Music: Learning to Generate Music from Videos”, ECCV, 2020.
4) Gunjan Aggarwal, Devi Parikh：“Dance2Music: Automatic Dance-driven Music Generation”, arXiv:2107.06252, 2021.