E-056

インタラクティブ性の高い音楽演出のための自然なトランジションのリアルタイム生成手法の検討

Exploring Real-Time Generation of Natural Transitions for Interactive Musical Performances

村瀬 朱音†
Akane Murase

1. はじめに

ビデオゲームにおいて BGM は臨場感や没入感を高めるために不可欠であり，近年ではプレイヤーの状況や操作に応じてリアルタイムに変化することが一般的である．ただ BGM を変化させるとは言っても，異なる楽曲・展開に移行する「横の遷移」では，ゲームの状況が変化した瞬間に BGM を唐突に切り替えたり，逆に音楽の繋がりを重視して BGM の切り替えが遅延したりすると，ゲーム体験が損なわれてしまうため，音楽を速やかにかつ自然に遷移させることが重要である．そのようなインタラクティブな音楽演出は既に多くの作品で採用されているものの，素材の準備や遷移タイミングのアノテーションなどに高いコストがかかるという課題がある[1]．この問題に対して，自然な横の遷移を低コストで実現するために，任意のタイミングで2つの楽曲の間を滑らかに接続する「トランジション」を自動生成する研究が進められている．しかし先行研究ではコード進行の生成やモノフォニックな旋律の補間が主に扱われており[2][3]，ポリフォニックな構造の音響全体の遷移，リアルタイム性についての考慮は不十分に見受けられる．そこで本研究では，複雑な音響構造のトランジションをリアルタイムで自動生成する手法について検討した．今回の提案手法では，画像生成モデルである StyleGAN2[4]を用いて，2つの楽曲の特徴を融合したメルスペクトログラムを生成し，トランジションに変換する．生成する画像を高品質にするためには拡散モデルの利用も考えられる．しかし本研究ではリアルタイム性の観点から GAN を用いることにした．本研究では実験として，ドラムループを素材としてトランジションを生成し，アンケートによって提案手法を評価した．その結果，提案手法の有効性が示唆され，また改善点が見出された．

2. 提案手法

ここでは，本研究で検討したトランジションの生成手法について説明する．まず，遷移前の音源(source)および遷移後の音源(target)に対応する潜在ベクトルを用意する．次に，2.1，2.2で述べる2種類の手法と，2.3で述べる2種類の遷移度合いを組み合わせた計4種類の方法で source と target の補間を行い，Hung ら[5]が looperman dataset から抽出した23,983件のドラムループを用いて訓練した StyleGAN2を使用して，補間結果のメルスペクトログラムを生成する．最後にボコーダとして Hung ら[5]が提供している MelGAN を用いて，メルスペクトログラムからオーディオを生成する．

2.1 入力ベクトルのモーフィング

補間手法の一つとして，StyleGAN2への入力ベクトルのモーフィングを行う．具体的には，source と target に対応する2つの潜在ベクトル z_1, z_2 を用いて，Niu ら[6]に倣って z_1, z_2 にモーフィングを適用したベクトルz_αを球面線形補間(Slerp)で求めて利用する．

$$\mathrm{z}_\alpha = slerp(z_1, z_2, \alpha) \ (0 \le \alpha \le 1)$$

2.2 スタイルミキシング

StyleGAN2では，入力の潜在ベクトルから中間潜在ベクトルを介してスタイル変換のパラメータを導き，それを画像生成において，解像度を上げつつ，各解像度の層に注入して利用する．スタイルミキシングとは，この構造を利用して，2つの画像から得られるスタイル変換のパラメータを併用し，両画像の属性を様々なレベルで融合させた画像を生成する手法である．この手法をメルスペクトログラムに適用すれば，低解像度の層ではリズムを，高解像度の層では音色を強く反映するなど，層ごとに異なる音響的特徴を制御可能と考えられる．そこで本研究では，低解像度層には target, 高解像度層には source に由来するスタイルを注入するスタイルミキシングをもう一つの補間手法として利用することにした．

2.3 遷移度合いの調整

本研究では，まずモーフィングの補間比率αを0.5に固定した単純モーフィング法(morp)，スタイルミキシングのスタイル切り替え位置を第4層に固定した単純ミキシング法(styl)で1小節分のトランジションを生成した．しかしこれらの方法では，source からトランジション，トランジションから target への移行時に音響的変化が急激になる傾向が見られた．そこで，より滑らかな遷移を実現するため，補間比率 $\alpha = 0.2, 0.4, 0.6, 0.8$においてそれぞれ4拍子1小節分のオーディオを生成し，これらを拍ごとに分割したうえで，各オーディオの1拍目，2拍目，…を順に接続することで1小節分のトランジションを構成する分割モーフィング法(morp-s)を導入する．スタイルミキシングに対しても同様に，スタイル切り替え位置を第2層から第5層まで順に変化させてそれぞれで1小節分のオーディオを生成し，拍単位に分割・再構成することで，より滑らかなスタイル遷移を実現する分割ミキシング法(styl-s)を導入し，1小節分のトランジションを生成した．

3. 評価実験

本研究では，提案した4通りのトランジション生成手法について，StyleGAN2で生成したドラムループを source, target とした実験を行い，アンケートで評価した．なお実

†京都大学大学院人間・環境学研究科
Graduate School of Human and Environmental Studies, Kyoto University

Copyright © 2025 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

験では，提案手法の比較対象として，トランジションの前半を source，後半を target として単純に接続する手法をベースライン(base)として加えた．以下では，実験と結果について説明する．

3.1 実験の設定

実験では，オーディオはすべて4分の4拍子でテンポ120とした．各オーディオの構成は，1小節ずつのドラムループである source，target の組に対して，同じく1小節のトランジションを生成し，それを source と target のループの間に挟み，全体で5小節とした．トランジションの位置は明確に認識されないことが望ましいため，ランダムに2〜4小節目のいずれかとした．このような設定のもと，今回はアンケートの回答時間も考慮して，source，target の組はランダムに6組生成して，それぞれの組に対して，提案手法と base の5つの手法でトランジションを生成し，合計でオーディオを30本用意した．なおトランジション1つ(2秒)の生成時間には最大で0.62秒を要した．アンケートでは，まず楽器演奏あるいは作曲経験の長さと，音楽鑑賞の頻度を尋ねた後，各オーディオに対して Cutajar の研究[3]を参考に2つの設問に回答を求めた．まず設問1では，遷移開始タイミングの認識のしやすさについて「1.明確にわかった」〜「5.全くわからなかった」の5段階で，設問2では遷移の滑らかさについて「1.唐突」〜「5.滑らか」の5段階での回答を求め，各選択肢を1〜5点のスコアで評価することにした．

3.2 結果

本研究では，3.1の実験を9名に対して実施した．そのうち1名については回答時間が極端に短かったため，当該の回答を除いた8件を分析の対象とした．表1にアンケートの評価結果を示す． 設問1では，表1に示したように，全体として大差ないものの， base が平均2.10点と最も低く，トランジションの開始が比較的知覚しやすかったと評価された．一方，最も高いスコアを示したのは styl-s で平均2.54点であった．楽器演奏あるいは作曲経験が1年以上の6名の回答に絞って集計したところ，設問1の平均点は総じて下がり，音楽に造詣のある被験者はトランジションの開始拍を明確に認識していることが見て取れた．設問2では表1に示すように，base が平均3.06点と最も低く，単純な接続では滑らかさに欠けることが明らかとなった．一方で，styl-s は平均3.79点, 標準偏差0.35と最も評価が高く，回答のばらつきも小さかった．音楽経験者6名の回答に絞ってスコアを計算したところ，styl-s を高く評価する傾向はより強まった．遷移度合いに注目して比較すると，両方の設問において morp よりも morp-s，styl よりも styl-s の方がスコアが高かった．このことから，拍単位で段階的に遷移させた構成が聴取上の自然さにつながったと考えられる．

4. 考察

まず，各手法のスコアに大きな差がつかなかったことについては，実験対象をドラムループに限定したことから，source と target の音響的特徴の差が比較的小さく，どの手法でも繋げやすい条件であったことが要因と考えられる．その上で，モーフィングの方がスタイルミキシングよりもスコアが低くなる傾向が見られた理由の一つとして，補間比率αと知覚的な source と target の混合比とが一致しているとは言えないことが考えられる．例えば，morp-s で提示

表 1　各手法（base：ベースライン法，morp：単純モーフィング法，morp-s：分割モーフィング法，styl：単純ミキシング法， styl-s：分割ミキシング法）におけるスコアの比較

	base	morp	morp-s	styl	styl-s
設問 1 平均	2.10	2.13	2.35	2.33	2.54
設問 1 標準偏差	0.42	0.35	0.35	0.58	0.42
設問 2 平均	3.06	3.33	3.58	3.63	3.79
設問 2 標準偏差	0.70	0.75	0.64	0.67	0.35

されたオーディオの中には，3段階目の補間比率である$\alpha = 0.6$においてもリズムがほぼ完全に source 寄りに聞こえるものがあった．このケースでは残りの拍($\alpha = 0.8$)だけでは滑らかに遷移しきれず, スコアが低く評価されたと考えられる．また，morp においても，$\alpha = 0.5$であるときのトランジションが必ずしも知覚的にちょうど中間として感じられるとは限らないため，知覚的な中間とのズレが大きかったケースがスコアを下げる一因となった可能性がある．

5. おわりに

本研究では，ゲーム音楽での横の遷移を想定して，StyleGAN2から作られる source，target を自然につなぐトランジションをリアルタイムに自動生成する手法を提案し，その有効性を検証した．実験ではドラムループを対象として，アンケート結果からは，本研究で提案した4つの手法のうちでは，styl-s が他の手法と比較して滑らかなトランジションが生成できる可能性が示唆された．一方で，本研究の実験で用いた音源は4分の4拍子，テンポ120のドラムループというごく単純な構成であったため，今後はより多様な音楽への対応が課題となる．また球面線形補間によるモーフィングにおいては，補間比率αと知覚される音楽の変化の度合いにズレが生じうることが実験で示唆された．この課題に対しては，Niu ら[6]が提案する音響知覚距離比率(SPDP)を導入することで，より自然な知覚的変化を実現できる可能性がある．提案手法の利便性を高めるためには，楽曲の潜在ベクトルへのエンコードも今後の課題である．

謝辞

この研究を進めるにあたり，貴重なご指導と温かいご助言を賜りました指導教員の日置尋久教授に心より感謝申し上げます．

参考文献

[1] 株式会社バンダイナムコスタジオ. 聴いて分かる！インタラクティブミュージック作曲の舞台裏！！. https://speakerdeck.com/bandainamcostudios/behind-the-scenes-of-interactive-music-and-composition. 最終アクセス 2025-06-01.
[2] Hadimlioglu, I. Alihan, and Scott A. King. "Automated musical transitions through rule-based synthesis using musical properties." *Entertainment Computing* 28 (2018): 59-67.
[3] Cutajar, Simon. *Automatic generation of dynamic musical transitions in computer games*. Diss. The Open University, 2020.
[4] Karras, Tero, et al. "Analyzing and improving the image quality of stylegan." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
[5] Hung, Tun-Min, et al. "A benchmarking initiative for audio-domain music generation using the freesound loop dataset." Proc. of the 22nd Int. Society for Music Information Retrieval Conf. (2021) 310-317.
[6] Niu, Xinlei, Jing Zhang, and Charles Patrick Martin. "SoundMorpher: Perceptually-Uniform Sound Morphing with Diffusion Model." *arXiv preprint arXiv:2410.02144* (2024).

Copyright © 2025 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.