令和元年度 卒業論文

動画の盛り上がり度に基づいた
ループシーケンサ

指導教員 北原鉄朗准教授

日本大学文理学部情報科学科

安坂文汰

2020年2月 提出

概　　要

今日，スマートフォンの普及により，Youtube やニコニコ動画などの動画共有サービスが人気を博している．これらの影響によって人々が動画を編集する機会が増加し，動画編集を支援するアプリケーションも普及した．より良い動画を作成するためには，バックグラウンドミュージック（BGM）の付与は必要不可欠だろう．しかしこのようなアプリケーションを用いても，使用する楽曲は自身で探す必要がある．

自動作曲のアプローチは様々なものがあり，動画を基に自動作曲を行う研究もある．動画から自動作曲では色や動きなどの動画特徴量と調やコードなどの音楽特徴量に対して適切な対応付けをすることで楽曲生成をするがこれは必ずしも簡単ではない．

本稿では既存研究で触れられていなかった動画の盛り上がりと楽曲の盛り上がりの対応付けを行うことで楽曲生成を行う．まず，動画の盛り上がり度の時間軌跡を取得し，その時間軌跡からテクノミュージックの自動生成を行う．動画の盛り上がり度の時間軌跡取得では動画のフレーム間での画素の動きを特徴量として抽出する．動画の対してモーションテンプレート解析を行い，動画の躍動の度合いを求め，盛り上がり度を推定する．推定した盛り上がり度をできるだけ再現するように，隠れマルコフモデル（HMM）を用いてループシーケンサに挿入する音素材を決定する．ただし，楽曲の始まりから終わりまで同じ HMM を用いると，テクノ特有の楽曲構成を再現できないため，テクノの楽曲構成を考慮して HMM のパラメータを使い分ける．

以上の処理を実現したシステムの実装を行い評価実験を行ったところ，動画の対応については 7 段階中 4.86，自然なテクノミュージック生成については 7 段階中 4.60 となり，楽曲構成を考慮しない手法と比較したところ，10 種類の動画に楽曲を付与する実験を行ったところ，楽曲が動画との盛り上がりと合致しているかは 7 曲，自然なテクノミュージックになったか 6 曲において，楽曲構成を考慮しない場合に比べて高評価が得られた．

目　　次

図 目 次

表 目 次

第1章　序　　論

本章では，研究の背景，目的を述べた後，本論文の構成を述べる．

1.1　本研究の背景

今日，スマートフォンの普及により，Youtube やニコニコ動画などの動画共有サービスが人気を博している．これらの影響によって人々が動画を編集する機会が増加し，動画編集を支援するアプリケーションも普及した．より良い動画を作成するためには，バックグラウンドミュージック（BGM）の付与は必要不可欠だろう．しかしこのようなアプリケーションを用いても，使用する楽曲は自身で探す必要がある．使える楽曲には，著作権などの問題によって限りがある．それに加えて，素人が作曲を行うのは難しいため，楽曲を用意するのは困難である．仮に見つかったとしても，動画の長さに合うように調整する，といった手間や技術が必要となる問題がある．

動画には，ストーリーやメッセージを伝えることを重視する場合の他に，躍動感，スピード感，スケール感，疾走感のような様々な印象を伝えることを重視する場合がある．一方，躍動感やスピード感を重視する音楽ジャンルにテクノミュージック（以下，テクノ）がある．テクノはクラブミュージックの一種であり，比較的単純なドラムパターンを何度も繰り返すことでスピード感を出し，次々に音素材を加えていくことで楽曲としての盛り上がりやメリハリを表現する．このような動画の躍動感とテクノの盛り上がりを対応付けて楽曲生成を行えば動画にあった楽曲を生成できると期待されるが，過去にはこのような研究はない．

1.2 本研究の目的

本研究では動画に合った楽曲を取得するために，動画の盛り上がり度に注目した楽曲の自動生成に取り組む．上で述べたように動画の盛り上がりと楽曲の盛り上がりを対応付けた楽曲生成は過去にはないため，動画の持つ盛り上がりからの楽曲の自動生成を行う．テクノの生成には飯島らのループシープケンサ [3] を拡張したものを用いる．ここでは動画の盛り上がりに依存しないように，テクノの基本的な楽曲構成を考慮できるように拡張する．

本稿では既存のループシーケンサ [3] と同じ手法を用いたシステム，楽曲の構成を考慮したシステム，楽曲の構成を考慮し，音素材の変化を減らしたシステムによって楽曲を生成を行う．生成された楽曲が動画の盛り上がりに対応できているか，楽曲として自然かどうかについて評価実験を行ったため報告する．

1.3 本論文の構成

本論文は次の構成からなる．第 2 章では，現状の自動作曲ではどのような手法が取られているのか，また動画を基に自動作曲に対して関連研究を紹介しながら述べる．第 3 章では，動画の盛り上がり度の推定，推定された盛り上がり度を反映した楽曲の生成方法を述べる．第 4 章では，本研究の使用環境や使用データの準備，実験方法と実験結果，そして考察について述べる．第 5 章では本研究の結論，また今後の課題について述べる．

第2章　関連研究

本章では関連する研究を紹介し，それらで未解決の課題を述べ，本研究における方針を述べる．

2.1　自動作曲に関する研究

自動作曲に関する研究では，音楽知識の少ない人でも簡単に作曲できるループシーケンサがある [3]．ユーザがマウスから入力した曲線を盛り上がり度に変換し，変換した盛り上がり度を基に挿入する音素材を HMM によって決定することでユーザの意志を楽曲に反映できる．しかし，動画に音楽を付ける際にはユーザが動画の盛り上がりを考えながら曲線を描画する必要がある．

その他に，旋律を音の経路と捉え作曲を経路探索問題として定式化し，任意の日本語の歌詞を用いた歌唱曲の自動作曲が，歌詞の韻律に基づく制約条件下での最尤経路探索問題を解くことで楽曲を生成するシステムがある [4]．このシステムではユーザが入力した歌詞から自動で楽曲を生成できるため，誰でも簡単に作曲を楽しむことができる．しかし，現在のシステムでは，動画からの自動作曲は難しい．

2.2　動画に対する楽曲付与に関する研究

2.2.1　動画解析・印象推定による動画 BGM の自動生成 [5]

動画から一定時間ごとに抽出した動きや色の動画特徴量から動画の印象を推定し，その結果に基づいて楽曲生成を行うことで，動画の印象に合った楽曲を付与するシステムである．ユーザに予め印象を回答してもらったリズム・メロディ素材をマッシュアップすることで楽曲生成を行うことから，ユーザごとの印象の違いを考慮した楽曲生成が可能となる．これにより，印象に合った音楽を自分で探すことなく動画に付与することができる．しかし，手軽に修正を行えない問題点がある．また，ここでは動画の盛り上がりは考慮されていない．

2.2.2　動画をもとにした自動作曲 [6]

動画の色情報に音楽心理学を対応付けることで，スケールやコードを決定することで作曲を行うシステムである．ここでは動画の中で画面が大きく変化するタイミングを「場面転換」としてブロック分けを行い，それぞれのブロックにおいて BGM を作成する．RGB 成分からコード進行の決定，HSB 成分からリズムを決定する．しかし，ここでは動画の躍動には触れておらず，盛り上がりを考慮していない．

2.2.3　映像の盛り上がり箇所に音楽のサビを同期させる BGM 付加支援手法 [7]

入力映像の指定箇所と入力楽曲の指定箇所を同期させながら，映像の全区間に対して BGM を付加する研究である．楽曲と映像の長さを揃えながら，ユーザが指定した楽曲と映像の箇所を同期させるように楽曲を断片的につなぎ合わせるこ

とで，映像の全区間に対して BGM を付加する. 具体的には, 動的計画法に基づく小節単位での楽曲の切り貼りによりユーザが指定した箇所を同期させた BGM の付加を実現するものである．この研究では動画の付与を行うものであり，BGM の生成は行っておらず，予め，音楽を用意する必要がある．

2.2.4 Generation of Personalized Music Sports Video Using Multimodal Cues[8]

スポーツビデオを分割部分を抽出し，ビデオと音楽の構成を考慮したデータベースから音楽を導入することでパーソナライズされた音楽スポーツビデオを半自動生成するシステムである．実験結果とユーザー評価は有望であり，このシステムで生成されたミュージックスポーツビデオは，プロが生成したものに匹敵することを示している．この研究では動画がスポーツ動画に限られ，音楽が半自動生成であるため，汎用性があまり高くないと言える．

2.3 本研究における方針

[3][4] のように自動作曲システムは簡単な操作性で音楽知識がなくても楽曲を生成できる研究が多く存在する．これらは音楽知識がなくても作曲することが可能である．[4] では歌詞の韻律に基づいて楽曲を生成するため，歌詞のある楽曲を生成する．そのため，動画の BGM に生成するためには動画にあった歌詞を考える手間が生まれる．[3] ではユーザが入力した曲線を楽曲の盛り上がりとして楽曲生成を行うため，動画の BGM を生成するためには動画が持つ盛り上がりをユーザが入力する必要がある．[5][6][7] のように動画への BGM 付与を行う研究は存在した．しかし，これらの研究では動画と楽曲の盛り上がりの対応については考慮されていない．そこで本研究では動画の盛り上がりを考慮した楽曲の生成を行う．

第3章　システム構成

本章では動画に対する盛り上がり度の推定，また推定された盛り上がり度を反映した楽曲の生成方法について述べる．

3.1　システム概要

本稿では，躍動感やスピード感を重視して作られた動画にテクノの BGM を付与するシステムを提案する．テクノを作曲するツールの一つにループシーケンサがある．ループシーケンサは，通常のシーケンサのように音符を入力するのではなく，数秒の音素材を組み合わせて作曲するツールである．しかし，多数の音素材から適切なものを選ぶのは必ずしも簡単ではない．そこで，飯島らは，盛り上がり度の時間軌跡をマウスなどで入力させ，それに基づいて自動で音素材を選ぶことで，手軽に作曲できるようにしたシステムを提案した．本稿で提案するシステムは，このシステムの拡張にあたり，動画中の躍動から盛り上がり度を求めて，その盛り上がり度から BGM を自動的に生成する．飯島らのループシーケンサは，盛り上がり度の時間軌跡をマウスで入力させ，その盛り上がり度に基づいて挿入する音素材の個数や種類を決めていた．本システムは，動画に含まれる躍動（動画中のオブジェクトや動画全体の動きの量）から盛り上がり度の時間軌跡を求め，そこからは飯島らのループシーケンサと同じ手法で楽曲を生成する．この手法には，動画は盛り上がるほど躍動が大きくなり，また，BGM においても盛り上がるほど音の数が増えたり，音が派手になる，という前提がある．

しかし，動画には動画の，テクノにはテクノの盛り上がり方があり，単に動画

中の躍動に忠実に音素材を挿入しても，テクノとして適切なものになるとは限らない．そこで，テクノに関わりの深い Electronic Dance Music (EDM) の典型的な楽曲構成を参考に，楽曲構成に制限を加えることで，動画にある程度合わせつつもテクノとしての適切さを失わないようにする．

以下，動画から盛り上がり度を求め，そこから BGM を生成する手法を述べる．現状の実装では，用いる音素材はすでに 1 小節単位で提供されており，すべて同じテンポであることを前提としている．つまり，1 小節あたりの長さは一定とする．楽曲の長さは 4 小節の倍数であることが多いことを考慮し，動画の長さを超えない範囲で，できるだけ楽曲が長く小節数が 4 の倍数になるように小節数を決定する．以下，小節数を N とする．

3.1.1　動画からの盛り上がり度の抽出

まず，BGM を付与したい動画を読み込む．読み込んだ動画のフレーム間に対してモーションテンプレート解析を行う．「モーションテンプレート」は MIT Media Lab が開発した動画の動き抽出の効率的な方法である [9] [10]． 本手法では，各フレームの画像を複数の正方形のきまった幅にブロックに分割し，ブロックごとに，隣り合うフレーム間の動きの有無を検出する．動きの検出のあったブロック数を合計し，そのフレームにおける盛り上がり度とする．

以降の処理は小節単位で行うため，各小節に属するフレームの盛り上がり度の平均値を，その小節の盛り上がり度とする．以降，各小節の盛り上がり度を $x_0, \cdots, x_{N-1}$ とする．この方法で求められる盛り上がり度は，動画によってその値の幅が大きく異なるため，次式で正規化して 5 段階に離散化する:

$$x'_n = \mathrm{round}\left(\frac{4x_n}{\max(x_0, \cdots, x_{N-1})}\right) \tag{3.1}$$

ここで，x'_n が離散化後の盛り上がり度，max は最大値を求める関数，round は

四捨五入を行う関数とする．

3.1.2 楽曲の構成の決定

上で述べたように，動画とテクノには本来異なるスタイルの盛り上がり方があり，単に動画の盛り上がりの軌跡に合わせて楽曲を生成するだけでは，テクノとして適切なものにはならない可能性がある．

Web ページ [1] によると，EDM は 4 個のセクションに分割することができる．楽曲の最初である「Intro」はペースを整え，ディスクジョッキー（DJ）が前の曲と繋ぐために使われるのが一般的である.「Intro」が終わると,「Breakdown」に移り，聴いている人へ期待感を与える.「Breakdown」は「Buildup」へつながり，徐々に盛り上がった後，メインである「Drop」へと遷移する．Drop は，楽曲の中で最大の盛り上がりとなる．一方，Web ページ [2] では，EDM は上記のセクションにおける「Drop」の後ろに，曲をつなぐための「Outro」を加えた 5 個のセクションから構成されると言われている．これらを考慮し，本稿では「Intro」,「Breakdown」,「Buildup」,「Drop」,「Outro」のセクションを使用する．

セクションの割当は 4 小節毎に行う．それぞれのセクションは想定されている盛り上がりが違うと考えられる．そのため，それぞれのセクションに対して標準的な盛り上がりを設定し，(3.1) 式で求めた盛り上がりとの差が最小になるように，セクションを割り当てる．なお「Buildup」は「Breakdown」の後ろから 2 小節に適用する．

前節の手法で求めた小節ごとの盛り上がり度を $x'_0, x'_1, \cdots, x'_{N-1}$ とする．いま，楽曲構成における各セクションの遷移は 4 小節単位で行われると仮定しているので，4 小節ごとの盛り上がり度を次式で定義する:

$$x''_n = \left(\frac{x'_{4n} + x'_{4n+1} + x'_{4n+2} + x'_{4n+3}}{4} \right) (n = 0, \cdots, N/4 - 1) \tag{3.2}$$

また，各セクションの標準的な盛り上がり度を $y_{\text{intro}}, y_{\text{breakdown}}, y_{\text{drop}}, y_{\text{outro}}$ とする．2 つの時系列 $\{x''_0, \cdots, x''_{N/4-1}\}$ と $\{y_{\text{intro}}, y_{\text{breakdown}}, y_{\text{drop}}, y_{\text{outro}}\}$ に対して dynamic time warping (DTW) を適用することで，セクションと小節の対応を決定する．なお，現在の実装では，$y_{\text{intro}} = 0.0$, $y_{\text{breakdown}} = 1.0$, $y_{\text{drop}} = 3.3$, $y_{\text{outro}} = 0.0$ としている．局所距離関数としては，単に差の絶対値を用いる．

3.1.3 音素材の挿入

飯島らのループシーケンサ [3] と同様の手法で音素材を挿入する．音素材として用いる「Sound PooL vol.1」[11] の仕様に基づき，音素材を「Sequence」,「Synth」,「Bass」,「Drums」の 4 パートに分け，各小節には各パート最大 1 つの音素材を挿入できるものとする．各音素材には 0~4 の 5 段階で盛り上がり度が付与されているとする．挿入する音素材の決定は次の 2 つのステップで行われる．

1. 挿入するかどうかの決定.
2. 挿入するなら何を挿入するかの決定.

挿入するかどうかの決定

各パートに音素材を挿入するか，しないかの 2 つの選択肢があるため，1 小節における音素材のパターンは 2^4 となり，これを状態とみなす．以降，各小節の状態を $s_0, \cdots, s_{N-1}$ とする．(3.1) 式で求めた盛り上がり度を出力信号とみなせば，隠れマルコフモデルにより解決することができる．初期確率（**表 3.1**）と状態遷移確率はセクション毎に設定し，状態遷移確率は初期確率をそれぞれの状態に設定した．出力確率（**表 3.2**）は，音素材が挿入されるパートが多い状態ほど，高い盛り上がり度に出力される様に設定した．例えば「Intro」では高い盛り上がりを観測しても，挿入する音素材の数は少なくなるように,「Drop」では観測された盛り上

がり度が低くても挿入する音素材の数は多くなるように遷移確率と初期確率を設定している．

挿入するなら何を挿入するかの決定

挿入する音素材は楽曲に一貫性をもたせるため，4 小節毎にそれぞれのパートと盛り上がり度に対してランダムで選択する．$x'_0, x'_1, \cdots, x'_{N-1}$ と $s_0, s_1, \cdots, s_{N-1}$ をもとに対応した音素材を挿入する．つまり，n 小節目の音素材は，次のように決定される．n が 4 で割り切れる場合，上で求めた状態 s_n において音素材が挿入されることになったパートは，盛り上がり度が x'_n の音素材からランダムに選択される．n が 4 で割り切れない場合，直前の小節の音素材がそのまま使われる．

表 3.1: セクション毎の初期確率. 表内の IT は Intro，BD は Breakdown，BU は Buildup，DR は Drop，OT は Outro を表す．状態は右から Drums，Bass，Synth，Sequence を表し，○の時挿入され，×の時挿入しないものとする．

状態	IT	BD	BU	DR	OT
××××	0.00	0.00	0.00	0.00	0.00
×××○	0.10	0.06	0.05	0.03	0.05
××○×	0.06	0.06	0.05	0.03	0.05
××○○	0.10	0.15	0.10	0.03	0.05
×○××	0.06	0.06	0.05	0.03	0.10
×○×○	0.06	0.06	0.05	0.03	0.05
×○○×	0.06	0.06	0.05	0.03	0.05
×○○○	0.06	0.10	0.20	0.03	0.05
○×××	0.06	0.06	0.05	0.03	0.15
○××○	0.06	0.06	0.05	0.03	0.05
○×○×	0.06	0.06	0.05	0.03	0.05
○×○○	0.06	0.06	0.05	0.10	0.05
○○××	0.06	0.06	0.05	0.03	0.15
○○×○	0.06	0.06	0.05	0.03	0.05
○○○×	0.06	0.06	0.05	0.20	0.05
○○○○	0.06	0.06	0.05	0.30	0.05

表 3.2: 全セクションへ適用する出力確率. 表の横軸の数値は盛り上がり度を表す. 状態は右から Drums，Bass，Synth，Sequence を表し，○の時挿入され，×の時挿入しないものとする.

状態	0	1	2	3	4
××××	0.00	0.00	0.00	0.00	0.00
×××○	0.40	0.20	0.20	0.10	0.10
××○×	0.39	0.21	0.20	0.10	0.10
××○○	0.25	0.30	0.25	0.10	0.10
×○××	0.39	0.21	0.20	0.10	0.10
×○×○	0.25	0.30	0.25	0.10	0.10
×○○×	0.25	0.30	0.25	0.10	0.10
×○○○	0.15	0.20	0.30	0.20	0.15
○×××	0.35	0.25	0.20	0.10	0.10
○××○	0.20	0.20	0.30	0.10	0.10
○×○×	0.20	0.20	0.35	0.15	0.10
○×○○	0.10	0.10	0.20	0.25	0.35
○○××	0.20	0.20	0.35	0.15	0.10
○○×○	0.10	0.10	0.20	0.25	0.35
○○○×	0.10	0.10	0.20	0.25	0.35
○○○○	0.05	0.05	0.25	0.25	0.40

第4章　実　　験

本章では，実験方法と実験結果について述べる．

4.1　実験方法

本稿で提案するシステムは python3 を用いて実装した．動画のモーションテンプレート解析には OpenCV を用い，音素材は前に述べたように「Sound PooL vol.1」[11] の素材を使用した．

実験用動画に対して，本システムと比較用システムそれぞれが楽曲を付与した動画を生成する．比較用システムは既存研究 [3] と同じくセクションを考慮しないものと，セクションは考慮するが，4 小節間で音素材を統一しないもの 2 つを使用し，それぞれを手法 1，手法 2 とする．セクションを考慮し，音素材を 4 小節間で固定する提案手法を手法 3 とする．使用する動画は Free Video clips[12] 上にある動画を実験用に 32 小節分の長さに調整したものである．

動画 1∼10[13][14][15][16][17][18][19][20][21][22] を 32 小節の長さに合うように再生速度を編集した．上で述べた 3 つの手法によって楽曲を生成し，動画の付与したものを web 上に公開し，評価をしてもらう．動画それぞれに対して 26 人∼50 人の評価者が集まった．

回答者には作曲経験，テクノを聴くかどうか，ループシーケンサの使用経験を事前に回答してもらう．作曲経験は，

1. 全くない

2. ほとんどない

3. 作曲を試したことはある

4. 趣味としてたまに作曲する

5. 日常的に作曲し，作品を公開，演奏している

テクノを聴くかどうかは,

1. 全く聴かない

2. ほとんど聴かない

3. たまに聴いている

4. 毎日ではないが，普段から聴いている

5. 毎日のように聴いている

ループシーケンサの使用経験は,

1. 全くない

2. ほとんどない

3. ループシーケンサを使用して作曲を試したことはある

4. 趣味としてループシーケンサを使用して作曲している

5. 日常的に，ループシーケンサを使用して作品を作り公開している

のそろぞれ5段階で回答してもらう．その後，3つ手法に対して次の2つ質問を行う.

Q1 動画の盛り上がりに対応しているか

Q2 テクノとして自然な楽曲か

それぞれの回答に対してそう思うかどうかを 7 段階で評価してもらう．加えて理由を記述してもらう．

4.2 生成例

動画 3 に対して手法 1，手法 2，手法 3 を用いて BGM を生成した結果をそれぞれ**図 4.1**[1]，**図 4.2**[2]，**図 4.3**[3]に示す．システム画面の横軸は時間軸を表し，1 ブロックが 1 小節である．縦軸は上部と下部に分かれており，上部は盛り上がり度，下部は挿入された音素材の状態を表している．

読み込んだ動画は人々が様々な場所を歩く様子を撮影したものであり，背景の動きが大きく盛り上がり度の推定に影響した．雑踏を歩く場面で高い盛り上がり度が見られ，逆に人の少ない場所や建物のみが映っている場面では低い盛り上がり度が見られた．

手法 1 では，推定された盛り上がり度が音素材の挿入される数がそのまま反映されるため，**図 4.1** の 13 小節目で，動画が突然動きの少ない場面に変化するため，音素材が大きく減少した．著者が聴いたところ，音素材の遷移に一貫性がなく，テクノとして不自然な生成となった．

手法 2 では，セクション分割を行い，それぞれに HMM を適用しているため，盛り上がり度の曲線とセクションの両方を考慮した音素材の挿入が行われた．手法 1 で問題だった 13 小節目ではセクションが「Drop」内であるため，HMM により音素材の数は減少するものの，大幅な減少ではなかった．著者が聴いたところ，音素材の状態数に大きな変化がないが，挿入される音素材に変化が多かった．これ

[1]動画 1：https://youtu.be/bQWnn3UKiIc
[2]動画 2：https://youtu.be/mADGwJl_IAk
[3]動画 3：https://youtu.be/iytF0hWKeYE

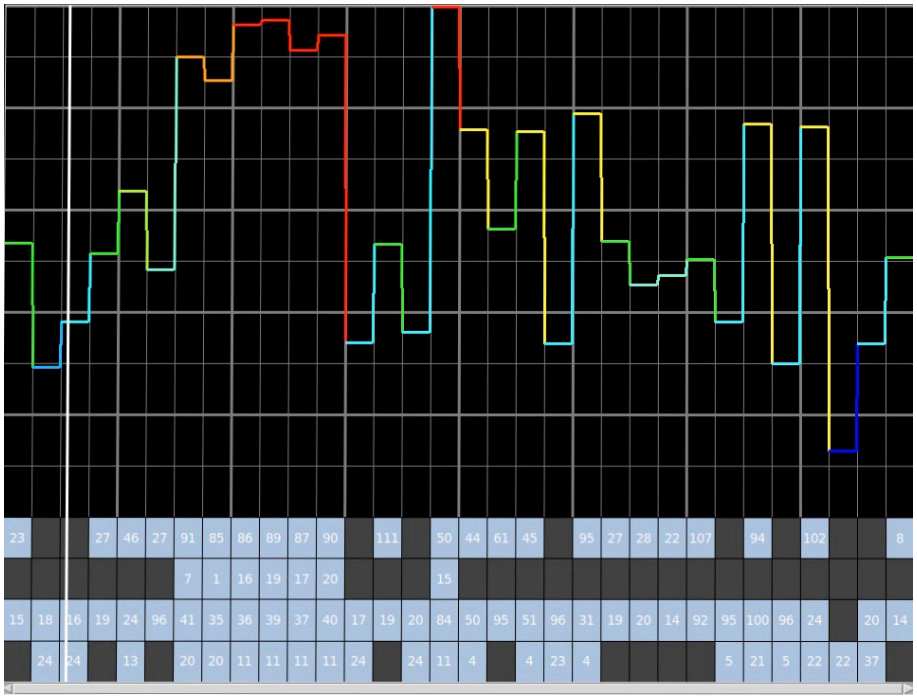

図 4.1: 手法 1 を用いた盛り上がり度推定と音素材の挿入

は，１小節ごとにランダムに音素材が変わるからである．そのため，音素材の繋がりが不自然であった．

手法 3 では，手法 2 に加えて 4 小節間で音素材を揃えているため，推定された盛り上がり度の反映は少なくなったものの，13 小節目のような明らかに盛り上がり度の変化のあるところでは音素材の状態数が変化していた．著者が聴いたところ，音素材の挿入される数と挿入される音素材の変化ともに一貫性があり，3 つの手法の中で一番自然なテクノであると考える．

今回提案した手法では，４小節ごとに１小節目の盛り上がり度に基づいて音素材の挿入を行ったが，4 小節ごとの平均値や最大値，最小値などを用いることもできる．何を用いるのが最も自然な生成が行えるかは，今後確かめていく必要がある．

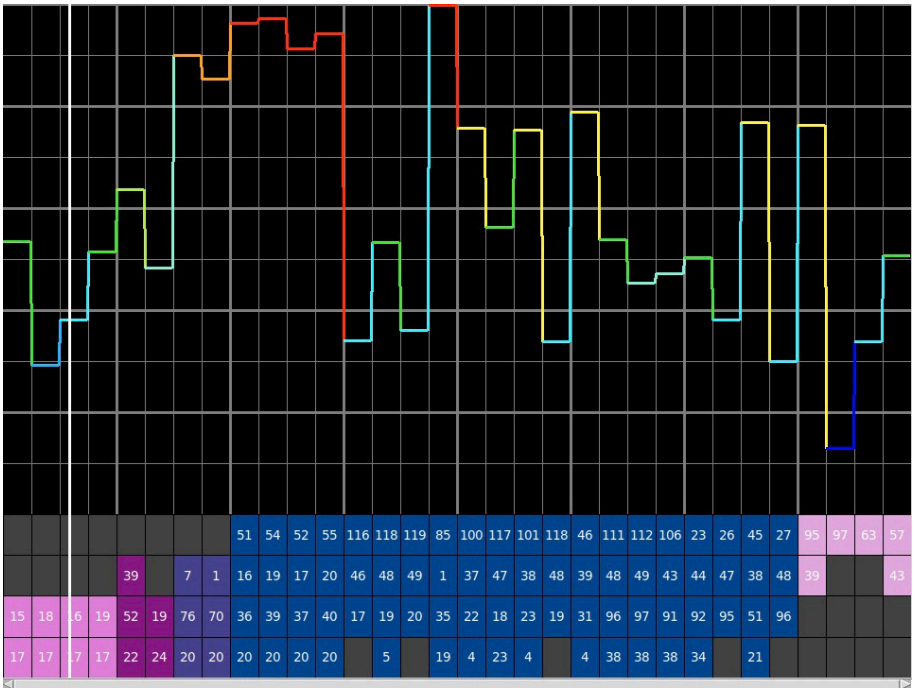

図 4.2: 手法 2 を用いた盛り上がり度推定と音素材の挿入

4.3 実験結果と考察

質問 Q1，Q2 に対する評価の平均値と標準偏差を図 4.4，図 4.5，図 4.6，図 4.7，図 4.8，図 4.9，図 4.10，図 4.11，図 4.12，図 4.13 に示す．

4.3.1 動画の盛り上がりに対応しているか（Q1）

評価結果を見ると，全ての動画における評価の平均が手法 1 は 3.93，手法 2 は 4.12，手法 3 は 4.30 となり手法 3 が最も高い値を取った．

動画別に結果を見ると，比較的大きな差が見られたのは動画 7 と動画 10 である．この 10 個の評価結果に対して t 検定を行った結果，動画 7 では手法 2 と手法 3 の間に，動画 10 では手法 1 と手法 3 の間で有意差が見られた．手法が 3 つあるため，多重検定の補正には Bonferroni 補正を用いた．結果を**表 4.1**, **表 4.2** に示す．

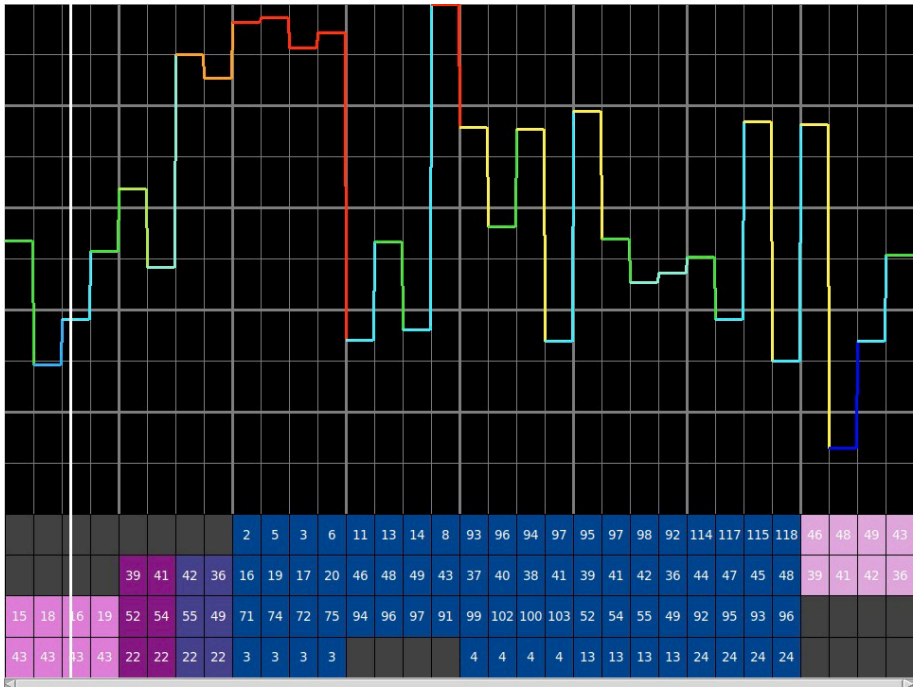

図 4.3: 手法 3 を用いた盛り上がり度推定と音素材の挿入

動画 7 は 1 つの滝を様々な角度からカメラで撮影する動画である．この動画では動画の中盤のカメラワークの変化が多い所で高い盛り上がりを抽出している．序盤と終盤では余り動きが見られずに低い盛り上がりを取り続けたため，動画の盛り上がりが比較的わかりやすいものであった．そのため生成する楽曲の構成に動画の盛り上がりがうまく反映できていたものだと考えられる．

動画 10 では人々の暮らしを映したものである．こちらの動画でも中盤のカメラワークの変化が多い所で高い盛り上がりを抽出している．動画 7 に比べて曲の変化は見られた．そのため動画 7 に同様，動画の盛り上がりが楽曲にうまく反映できたと考えられる．

今回の評価実験では動画のジャンルは問わなかったが，動画 7 と動画 10，この 2 つの動画から本手法の盛り上がり推定ではカメラワークの動きに大きく依存することがわかる．本手法へ対する評価者からのコメントには「カメラの動きには対応しているのかもしれないが，滝のイメージは湧いてこない」,「動画とは関係な

表 4.1: 動画 7 の Q1 に対する t 検定の結果（有意水準：0.05）

比較する手法	手法 1，手法 2	手法 2，手法 3	手法 1，手法 3
P 値	0.3510697661	0.006157587777	0.5418599775

表 4.2: 動画 10 の Q1 に対する t 検定の結果（有意水準：0.05）

比較する手法	手法 1，手法 2	手法 2，手法 3	手法 1，手法 3
P 値	0.5097223487	0.4899474906	0.02854064213

く音楽だけ盛り上がっている感じがする」などといった意見が見られた．本手法では動画中の動きの大きさのみを手がかりにしており，動画中に映っている物体が何かは考慮していないため，実際に映っている物体のイメージを表現できていないことが明らかになった．一方,「場面の転換に合わせて曲調が変わったから」,「画面に切り替わりと音楽が合っていると感じたため」などと肯定的なコメントが見られたことからカメラワークの動きは動画の盛り上がりには重要だと考えることができる．

10 個の動画の中でも動画 4 の評価が低かった．この動画も動画 10 に似たもので,人々の様子を映したものである．動画 10 に比べて画面内に多くに人が映っておりその 1 人 1 人の動きを抽出したため，フレーム内に映る人数が多いほど盛り上がり度が高くなる傾向がみられた．本手法へ対する評価者のコメントには「冒頭の犬が出てくるところなどは合っていたと思いますし，全体的に上げ下げの配分が良いように思いました.」,「盛り上がるような場面がないのに，曲が盛り上がってるのはおかしいと思うので,」など対立した意見が見られたため，動画の盛り上がりを感じる部分に個人差があるとも考えられる．

4.3.2 テクノとして自然な楽曲か（Q2）

評価結果を見ると，全ての動画における評価の平均が手法1は4.38，手法2は4.58，手法3は4.60であった．手法3が高い値を取った．手法2と手法3の差は大きくは見られなかった．

動画別に結果を見ると，比較的大きな差が見られたのは動画2，動画3，動画8，動画10である．Q1同様にt検定を行った結果，動画2では手法1と手法2，手法1と手法3の間で，動画3では手法2と手法3の間で，動画8では手法1と手法3の間で，動画10では手法1と手法2，手法1と手法3の間で有意差が見られた．結果を**表4.3**，**表4.4**，**表4.5**，**表4.6**に示す．

動画2は夜の町での人々を映した動画で場面が何度か切り替わる動画である．盛り上がりは中盤に高い盛り上がりを推定していた．動画の序盤と終盤は比較的動画の動きが少ないため，楽曲の構成が組みやすく，他の動画に比べて音素材の変化が自然になったと考えられる．本手法へ対する評価者のコメントには「音と音の間がスムーズに移動できたと感じるため」,「場面に自然と入り込んできます」などといった意見が見られ，他の手法に比べて自然な楽曲を生成出来ていることがわかる．

動画3は人がスケートボードで滑っている様子を映したものである．緩急をつけて滑っていたため，動画内の盛り上がるところとそうでないところの判断がしやすい動画だと言える．本手法へ対する評価者のコメントには「個々の部分で展開があるためか，手法1，手法2よりは音楽的に聞こえた」,「動画と合っていたかといえば微妙なところだけどテクノとしては自然な形になっていたと思います」と自然な楽曲への生成には肯定的なコメントが見られたが，動画との対応には不満を感じている評価者も見られた．

動画8は人々が電車内でダンスを繰り広げる動画である．ダンスは動きが激しい部分が多く，推定した盛り上がりは序盤から高い値を取り続けた．本手法に対する評価者のコメントでは「テクノミュージック特有の怪し気で不穏な雰囲気を感

表 4.3: 動画 2 の Q2 に対する t 検定の結果（有意水準：0.05）

比較する手法	手法 1，手法 2	手法 2，手法 3	手法 1，手法 3
P 値	0.005432417385	0.3582533275	0.0003401556018

表 4.4: 動画 3 の Q2 に対する t 検定の結果（有意水準：0.05）

比較する手法	手法 1，手法 2	手法 2，手法 3	手法 1，手法 3
P 値	0.4253132689	0.02780973408	1.384496603

じることが出来たことと，高音のメロディが入ってきても低音部分や怪しさが自然とキープされていたので，全体的にバランスの整った楽曲だと感じました」,「テクノミュージックとして違和感なく，聞いていて心地よい楽曲だった」といった意見が見られた．ダンス自体がテクノにマッチしていることためだと考えられる．

動画 10 では本手法に対する評価者にコメントに「概ね違和感なく，ひとつの楽曲として聞くことができたから」,「違和感も感じなかったし，自然な感じだったので」といった意見が見られ，手法 1，手法 2 に比べて自然なテクノを生成出来ていたと考える．

4 つの動画から本手法によって過去のシステム [3] に比べて自然なテクノを生成でにつなげる拡張ができたと言える．音の繋がりやまとまりへのコメントが多かったことから，音素材の挿入状態，入れる音素材の変化が良い結果へと繋がった．しかしコメントの中には「素人考えでは若干テクノではないように思う」,「あまりテクノミュージックを聞かないのですが，テクノっぽく感じる部分と感じない部分があるなという印象だったため」といった意見があったことからテクノの定義に個人差が見られた．そして「自然と言えば自然だけど，何かが物足りない感じ」,「単調なため，楽曲という印象はなかった」,「バスドラムの音域との被りが多すぎるため耳が疲れる可能性があります」,「全体的に不調和な気がした」と行った意見が見られたことから音域や調和を考慮した音素材の配置が課題となった．

表 4.5: 動画 8 の Q2 に対する t 検定の結果（有意水準：0.05)

比較する手法	手法 1，手法 2	手法 2，手法 3	手法 1，手法 3
P 値	0.2443120546	0.7109259992	0.02214228975

表 4.6: 動画 10 の Q2 に対する t 検定の結果（有意水準：0.05）

比較する手法	手法 1，手法 2	手法 2，手法 3	手法 1，手法 3
P 値	0.006586163812	1.186944663	0.001447922382

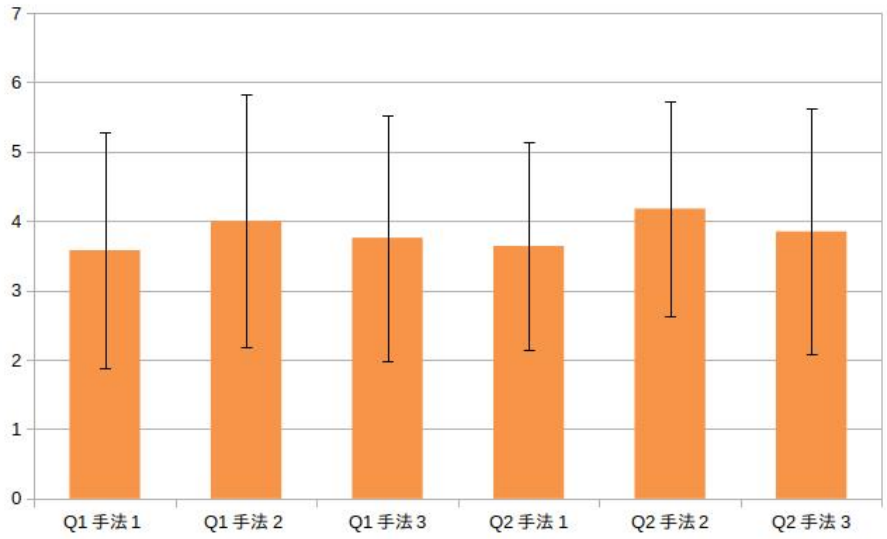

図 4.4: 動画 1 に対する質問の回答（回答者 33 人）

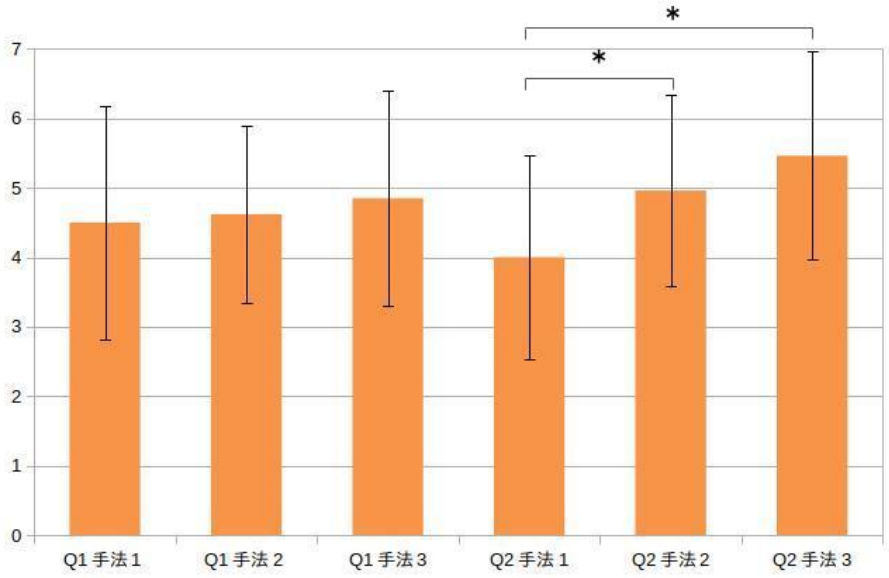

図 4.5: 動画 2 に対する質問の回答（26 人，$*$ $p < 0.05$）

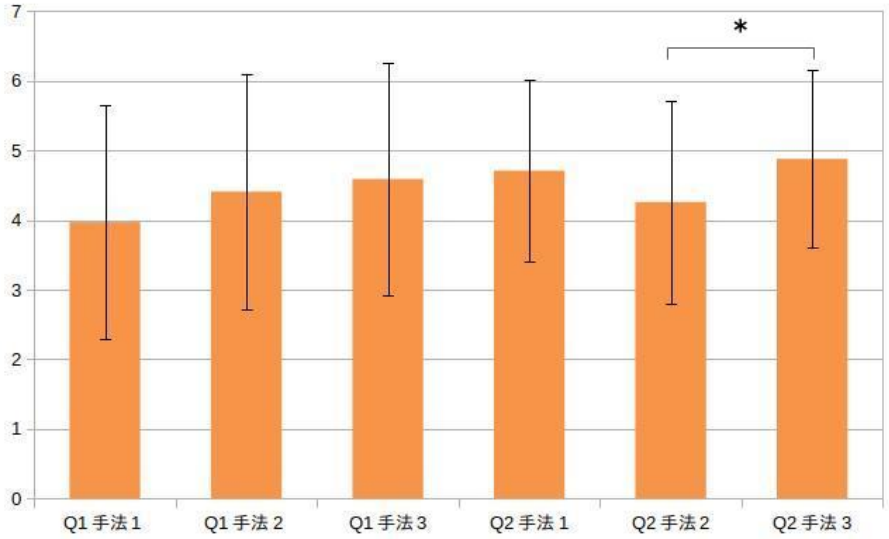

図 4.6: 動画 3 に対する質問の回答（34 人，$*$ $p < 0.05$）

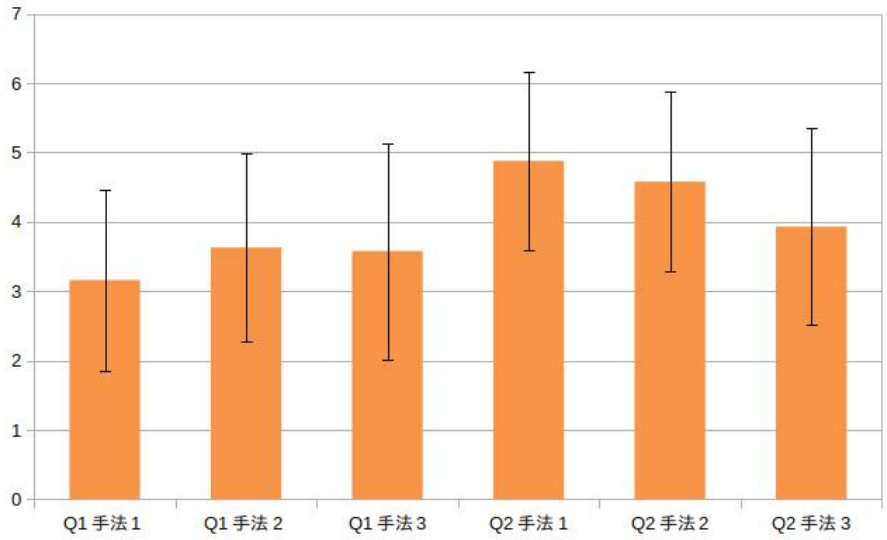

図 4.7: 動画 4 に対する質問の回答（43 人）

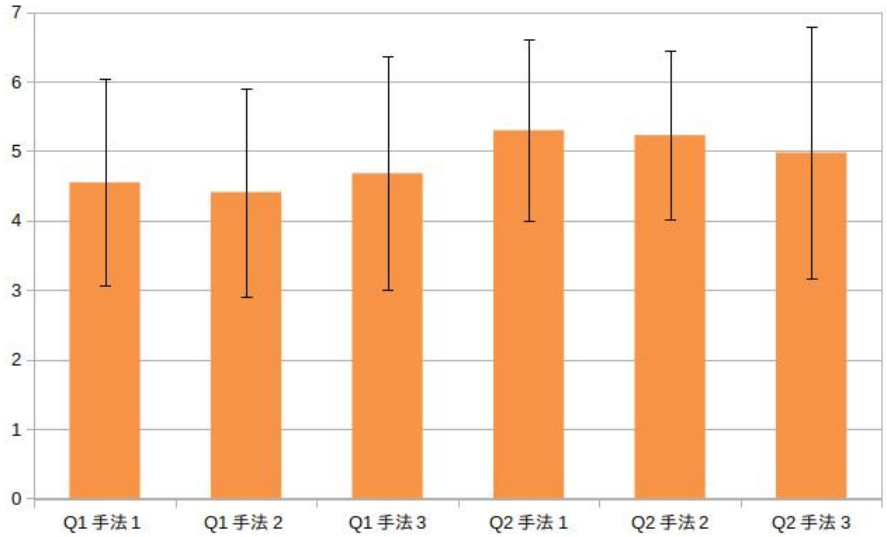

図 4.8: 動画 5 に対する質問の回答（44 人）

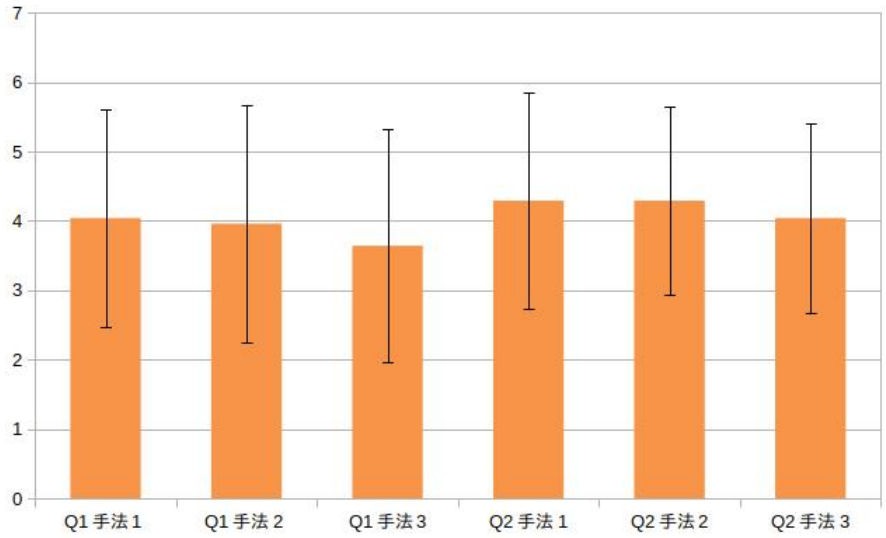

図 4.9: 動画 6 に対する質問の回答（28 人）

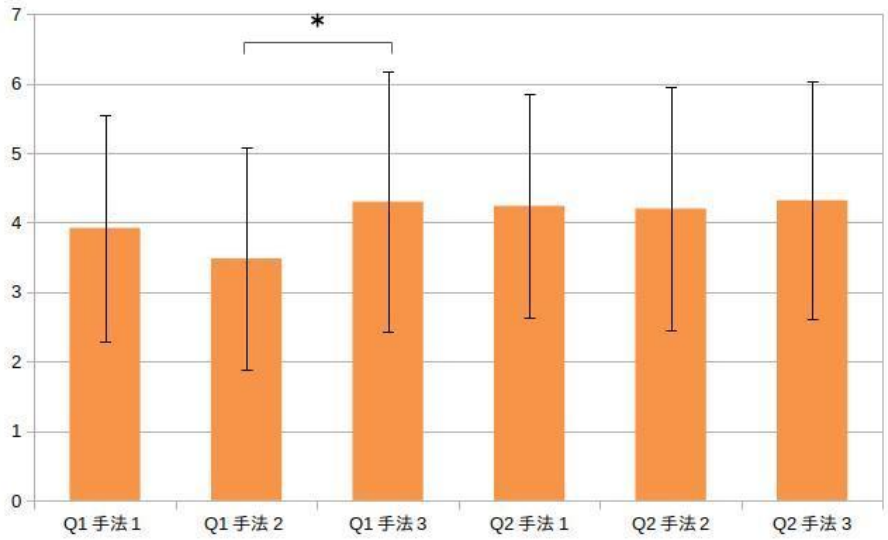

図 4.10: 動画 7 に対する質問の回答（50 人，$* \mathrm{p} < 0.05$）

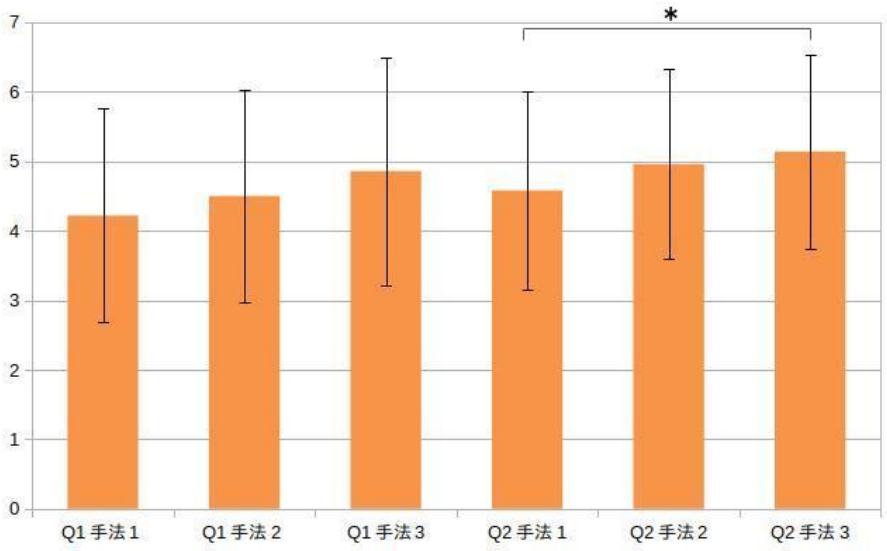

図 4.11: 動画 8 に対する質問の回答（50 人，$*$ $p < 0.05$）

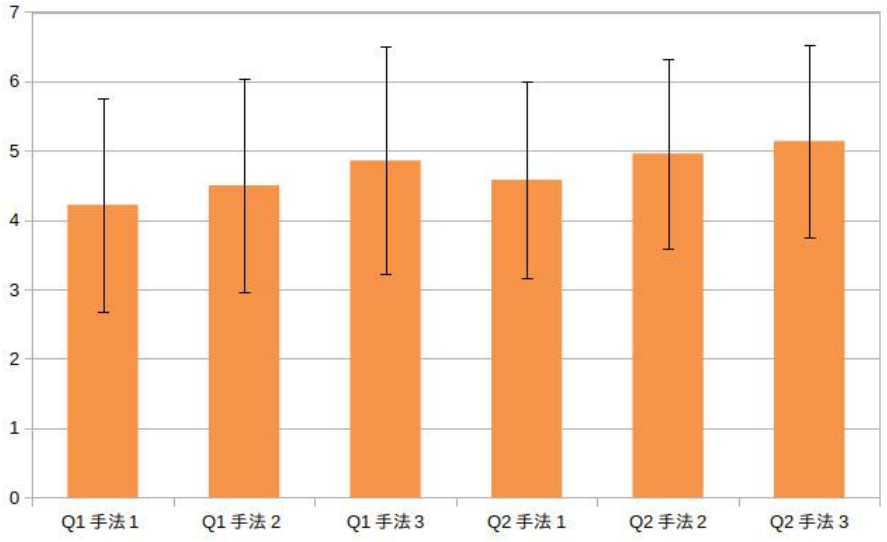

図 4.12: 動画 9 に対する質問の回答（33 人）

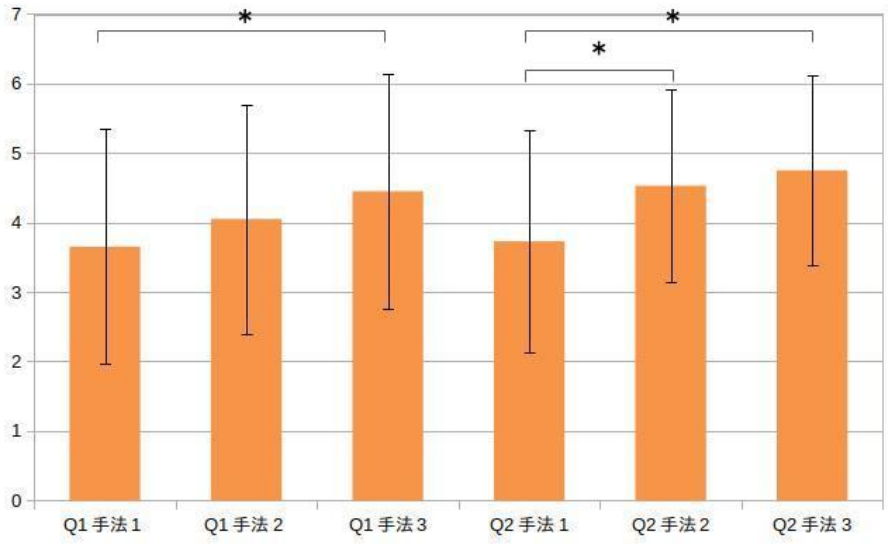

図 4.13: 動画 10 に対する質問の回答（40 人，$*$ $p < 0.05$）

第5章　結　　論

5.1　結論

飯島らが開発したループシーケンサ [3] は，描画した線から音素材を挿入するかどうか，挿入するならどの音素材を挿入するのかを HMM によって決定するため，直感的に作曲することができ，専門的な知識を必要としないため，誰でも簡単に作曲を楽しむことができる．しかし，現在のシステムでは入力が描画された曲線に限られ，作成される楽曲の構成も考慮されていない．そのため，動画を入力とした楽曲生成，構成を考慮した楽曲生成ができなかった．本稿では，動画の動きを特徴量とした盛り上がり度を推定し，楽曲の構成を自動で割り当てることで動画の盛り上がりに対応した楽曲を自動で生成する手法を提案した．実際に動画を読み込み，楽曲生成を行ったところ，楽曲構成を考慮した場合，音素材の挿入数が大きく増減しない遷移をする楽曲が生成された．さらに 4 小節毎に音素材を揃えることで音楽の変化を減らすことにより自然なテクノが生成を試した．本手法よって生成された楽曲を実際に著者が聴いたところ，既存研究より自然なテクノが生成できていた．

また，10 種類の動画に楽曲を付与し評価をしてもらう実験を行ったところ，評価者が 26 人 ~50 人集まった．結果は楽曲が動画との盛り上がりと対応しているかどうかは 7 曲，自然なテクノ生成ができたかどうかは 6 曲において，楽曲構成を考慮しない場合に比べて高評価が得られた．幾つかの動画では有意差を得ることができ，少なくとも，既存のループシーケンサ [3] と同じ手法を用いたシステムに比べて本稿で提案したシステムが有用なことを示せた．

5.2 今後の展望

今後，本システムの評価を回答者を増やして行うとともに，音素材の種類を増やし，様々な動画に対応した楽曲の自動生成を行いたい．今回はテクノに限定したが，その他の音楽ジャンルを持つ音素材にそれぞれ盛り上がり度をもたせることでテクノ以外の音楽を生成できるような拡張を予定している．その他，今回使用しなかった動画の色情報を分析することで，動画の盛り上がり度の推定方法の改善し，盛り上がり度の数を増やし，パート毎に挿入できる音素材の数を増やすことで，より多彩な楽曲を生成できるよう手法やシステムを拡張していきたい．

また，本システムをWebアプリケーションやスマートフォンアプリケーション上で実装し，一般の方々に広く使ってもらい，一般の方々の使い方を収集することでさらなるシステムの改良に活かしたい．また，本システムは単に動画の盛り上がり度をループシーケンサに入力するだけでなく，ユーザ自らの手で修正が可能であるため，修正を行うことで，楽曲をより自分好みに作り変えることができる．本研究のループシーケンサは盛り上がり度をマウスで描画するだけで使うことができるので，本システムを通じて，音楽の非専門家が自分好みの楽曲を自らの手でつくる文化を広めていきたい．

参考文献

[1] EDM Song Structure：Turn Your Loop Into A Song!-Cymatics.fm, https://cymatics.fm/blogs/production/edm-song-structure

[2] EDM の構成と作り方 | Madison Mars Milky Way(online), https://salondemuze.com/blog/tip/edm-structure/

[3] 飯島孔右, 鶴岡亜矢佳：手書き入力で盛り上がりをコントロールするループシーケンサ：スペクトログラムから盛り上がりの自動割り振り, 第 77 回全国大会講演論文集, Vol.2015, No.1, pp.373–374, 2015.

[4] 深山覚, 中妻啓, 米林裕一郎, 酒向慎司, 西本卓也, 小野順貴, 嵯峨山茂樹：Orpheus：歌詞の韻律に基づいた自動作曲システム, 情報処理学会研究報告音楽情報科学 (MUS), Vol.2008, No.78 (2008-MUS-076), pp.179–184, 2008.

[5] 清水柚里奈, 菅野沙也, 伊藤貴之, 嵯峨山茂樹, 高塚正浩：動画解析・印象推定による動画 BGM の自動生成, 研究報告音楽情報科学 (MUS), Vol.2015, No.17, pp.1–6, 2015.

[6] 山本敏生，宝珍輝尚，野宮浩輝，動画をもとにした自動作曲，平成 21 年度情報処理学会関西支部大会論文集, vol.2009, 2009.

[7] 佐藤晴紀，平井辰典，中野倫靖，後藤真孝，森島繁生，映像の盛り上がり箇所に音楽のサビを同期させる BGM 付加支援手法, 研究報告エンタテインメントコンピューティング (EC), Vol.2015, No.10, pp.1–6, 2015.

[8] J. Wang, E. Chng, C. Xu, H. Lu, Q. Tian: Generation of Personalized Music Sports Video Using Multimodal Cues, IEEE Transactions on Multimedia, Vol.9, No.3, pp.576–588, 2007.

[9] Davis, James W and Bobick, Aaron F：The representation and recognition of action using temporal templates, IEEE conference on computer vision and pattern recognition, pp.928–934, 1997.

[10] Davis, James and Bradski, Gary：Real-time Motion Template Gradients using Intel CVLib, IEEE ICCV Workshop on Framerate Vision, pp.1–20, 1999.

[11] Sound PooL | 製品情報 | AHS(AH-Software)(online),
https://www.ah-soft.com/soundpool/

[12] Free video clips(online),
https://mazwai.com/

[13] Free video clips - 9th May & Fireworks,
https://mazwai.com/video/9th-may-&-fireworks/455089

[14] Free video clips - never sleep,
https://mazwai.com/video/never-sleep/455029

[15] Free video clips - black and wine,
https://mazwai.com/video/black-and-wine/455046

[16] Free video clips - breathing barcelona,
https://mazwai.com/video/breathing-barcelona/455069

[17] Free video clips - No coast drift,
https://mazwai.com/video/no-coast-drift/454974

[18] Free video clips - Justin Wagers — PRECISION,
https://mazwai.com/video/justin-wagers-%7C-precision/455023

[19] Free video clips - La chute Delaney,
https://mazwai.com/video/la-chute-delaney/455085

[20] Free video clips - Between 14th & Bedford - NY Subway Dancers,
https://mazwai.com/video/between-14th-&
-bedford---ny-subway-dancers/455005

[21] Free video clips - In peace ,
https://mazwai.com/video/in-peace/454994

[22] Free video clips - slomosf,
https://mazwai.com/video/slomosf/455061

謝　　辞

本研究を進めるにあたり，北原鉄朗准教授から丁寧かつ熱心なご指導を受け賜りました．ここに感謝の意を表します．また，評価実験の際に評価者を快く引き受けてくださいました皆様，切磋琢磨し研究に励んだ北原研究室の同期，後輩に感謝いたします．