Vol.2015-MUS-106 No.10
Vol.2015-EC-35 No.10
2015/3/2

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

佐藤 晴紀[1,4,a)]　平井 辰典[1]　中野 倫靖[2]　後藤 真孝[2]　森島 繁生[3,4]

概要：本稿では，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を付加する手法を提案する. 従来研究では，既存動画から音響特徴量と映像特徴量を学習し，映像にBGMを自動付加する手法が提案されている. しかし，自動で映像にBGMを付加しているため，「映像の決定的なシーンに楽曲のサビを合わせたい」「映像の始端と終端に楽曲の始端と終端を合わせたい」といったユーザが指定した楽曲と映像の特定の箇所を同期するためのBGM付加については言及されていない. そこで本研究では，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. 被験者実験の結果，本手法は同じ音色の箇所が多いインストゥルメンタルの楽曲に対して特に有効であった．また，一度生成されたBGMをユーザが希望する楽曲の盛り上がりに合わせて再編集を行うことができるシステムを提案した.

1. 研究背景

近年，動画共有サイトの利用者の増加に伴い, 動画を共有する文化が広がっている．例えば，動画共有サイトYouTubeでは2015年1月の時点で，1分間に約100時間分の動画がアップロードされている．また，優れた無料動画編集ソフトの普及により動画編集に対する敷居が下がっており，個人が動画制作に携わる機会が増加している．さらに，動画編集ソフトの操作方法や効果的な色味の補正方法や音楽の使い方といった，動画制作に関するノウハウがWebを介して容易に入手することができる．これにより，動画制作の経験が浅いユーザも質の高い動画を制作することに関心が向けられるようになってきている.

動画制作における重要な工程にBGMの付加がある.BGMの付加は，映像の情景と合わせることで視聴者の映像に対する印象を強くさせる効果を持つなど，動画制作には欠かせない工程である [1]. また，BGMの付加により映像をより印象的に見せるために，編集者が「映像の決定的なシーンに楽曲のサビを合わせる」,「映像と楽曲の始まりと終わりを合わせる」などといったこだわりを持ったBGMの付加が行われている. さらに，映像の盛り上がりとBGM自信の盛り上がりを一致させるため「BGMの前半では盛り上がりを抑えて，後半で盛り上げたい」といった楽曲全体の盛り上がりを考慮しながらBGMを付加することがある. このようなこだわりや楽曲の盛り上がりを反映させるために，編集者は楽曲を切り貼りすることによってBGMを編集することがある. しかし，映像とBGMの長さを一致させるように調節させながら，切り貼りしたBGMの繋ぎ目に違和感が残らないようにするためには，楽曲のリズムや音色を繰り返し聴取しながら編集するといった，多くの手間暇がかかる反復作業が必要となる．さらに，楽曲の盛り上がりも同時に考慮してBGMを付加するにはより多くの労力が必要となる. このように，無数の繋ぎ方が存在するBGMの切り貼りのパターンの中から，いくつかの制約を考慮した上で最適な繋ぎ方を見つけることは困難である.

本稿では，ユーザが映像と楽曲を同期させる箇所を指定することで，映像と楽曲の始端と終端を合わせながら，映像と楽曲の指定箇所を同期させたBGMを自動生成する手法を提案する. 具体的には，動的計画法を用いた楽曲の小節単位での切り貼りによって，ユーザの指定箇所以外を補間することでBGMを生成する. しかし，生成されたBGMが必ずしもユーザの意図した盛り上がりと一致するとは限らない．そこで，一度生成されたBGMに対して，盛り上がりに関するユーザの意図を入力することで，それを反映

1 早稲田大学
Waseda University, Shinjuku, Tokyo 169-8555, Japan
2 産業技術総合研究所
3 早稲田大学理工学術院総合研究所
4 JST CREST
a) ha-ru-ki@asagi.waseda.jp

Vol.2015-MUS-106 No.10
Vol.2015-EC-35 No.10
2015/3/2

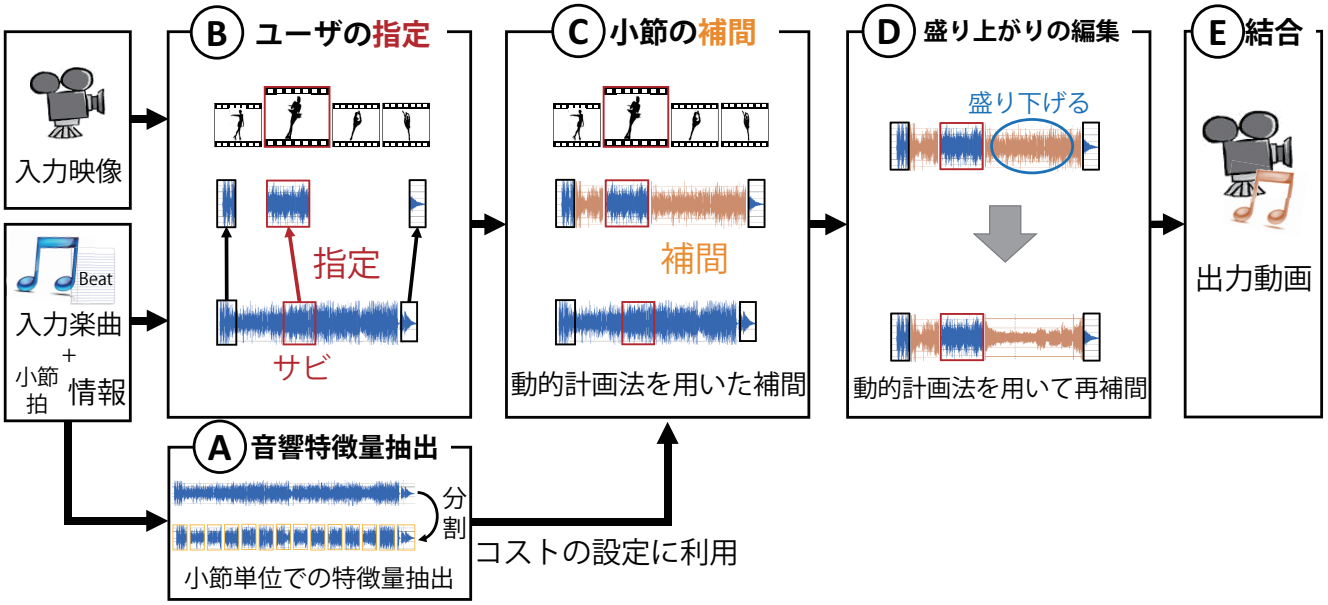

図 1　提案手法の流れ

するように BGM を再び生成し映像へ付加するシステムを提案する. それにより，楽曲の編集を経験したことのないユーザでも手間暇をかけずに，こだわりや楽曲の盛り上がりを反映した BGM の付加を可能とする．また，本手法が有効な楽曲の特徴を検証する実験を行った.

2. 関連研究

映像に BGM を付加する手法に関して，様々な先行研究がある．例えば，Yoon らは映像の画像特徴量と楽曲の音響特徴量を対応付けることで映像に BGM を自動付加する手法を提案した [2]．また，楽曲のムード分類を利用してユーザが映像の雰囲気に合った楽曲を選択し，選択された楽曲を BGM として自動付加するシステム [3] や，既存動画から映像と楽曲の関係を学習することで入力映像に対してデータベース内の楽曲を用いて BGM を自動付加する手法 [4] が提案されている．また，ダンス映像の被写体の動きやスポーツ動画の歓声に楽曲の音響特徴量を対応付けることで被写体の動きや，スポーツ動画のシーンに合った BGM を自動付加する手法が提案されている [5], [6], [7]．これらの既存手法では入力動画に合った BGM をデータベースの楽曲を用いて自動的に付加することができるが，単一楽曲の切り貼りを用いて，動画の特定のタイミングに楽曲の特定の箇所を同期させるユーザのこだわりや楽曲の盛り上がりを反映した BGM の付加については言及されておらず，ユーザの意図を反映させた BGM の付加を実現する手法は提案されていない.

また，楽曲の盛り上がりの算出に関しても手法が提案されている [8]，[9]．具体的には，楽曲のエネルギーと音楽構造を用いて楽曲の盛り上がりを算出する手法や楽曲のメロディラインとその音域のエネルギーに着目した手法がある.

3. BGM 生成及び付加の概要

本稿では，映像と楽曲の始端と終端を合わせながら，映像と楽曲の指定箇所が同期した BGM を生成する手法を提案する．本研究における提案手法の流れを図 **1** に示す．小節単位で楽曲を切り貼りする際，小節の繋ぎ目の音色が大きく異なるといった違和感が生じてしまうと，生成される BGM の質が下がってしまう. そこで，小節の繋ぎ目を自然にするために，境界付近が音響的に類似した小節同士を接続させる. そのため，入力楽曲の小節ごとに音響特徴量を抽出する（図 1, Ⓐ）. 次に，映像の始端と終端に楽曲の第一小節と最終小節を自動で配置し，映像と楽曲の指定箇所をユーザが入力する.（図 1, Ⓑ）. ここで，映像と楽曲の指定箇所の組は複数でもよい．そして，楽曲の第一小節の終わりと指定箇所の始まり，指定箇所の終わりと楽曲の最終小節の始まりの間を，動的計画法を用いた小節単位での楽曲の切り貼りによって補間することで，BGM を生成する（図 1, Ⓒ）. さらに，一度生成した BGM に対し，ユーザが好みの楽曲の盛り上がりを指定することで，ユーザの好みを反映させた新たな BGM を生成する（図 1, Ⓓ）. 最後に，補間する際に生じる映像と楽曲の指定箇所の微小な時間的ずれを補正し，映像と BGM を結合させることで，映像と楽曲の長さを一致させ，映像と楽曲の指定箇所を同期させながら，楽曲の盛り上がりを反映した BGM を映像へ付加する（図 1, Ⓔ）.

4. BGM 付加手法

4.1 小節の補間

楽曲の第一小節と指定箇所，最終小節以外の区間を動的計画法により補間する. 図 **2** に小節の補間の概要を示す.

Vol.2015-MUS-106 No.10
Vol.2015-EC-35 No.10
2015/3/2

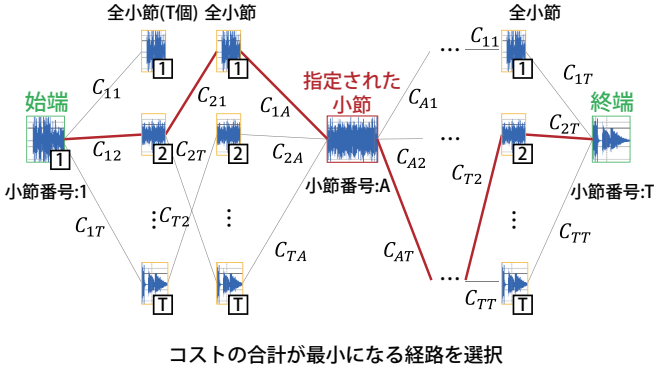

図 2　小節の補間

小節を補間する際，ある小節の次に繋がる小節の候補は楽曲中の全小節とする．つまり，小節同士のあらゆる繋がり方を考慮した上で動的計画法による経路選択を行う．j 番目の小節と k 番目の小節の繋がりやすさは式（1）に示すコスト関数に基づいて算出する．

$$C_{jk} = \alpha\sqrt{M + \Delta M + M_E + \overline{M} + M_\sigma} \tag{1}$$

$$\alpha = \begin{cases} 0 & if \quad k = j+1 \\ 1 & \text{else} \end{cases} \tag{2}$$

$$M = \sum_{t=1}^{N} ||{}^k\boldsymbol{m}_t - \boldsymbol{m}_t^j||^2 \tag{3}$$

$$\Delta M = \sum_{t=1}^{N} ||{}^k\Delta\boldsymbol{m}_t - \Delta\boldsymbol{m}_t^j||^2 \tag{4}$$

$$M_E = \sum_{t=1}^{N} \left({}^kE_t - E_t^j\right)^2 \tag{5}$$

$$\overline{M} = ||{}^k\overline{\boldsymbol{m}} - \overline{\boldsymbol{m}}^j||^2 \tag{6}$$

$$M_\sigma = ||{}^k\boldsymbol{\sigma} - \boldsymbol{\sigma}^j||^2 \tag{7}$$

$\boldsymbol{m}$ は MFCC（低 12 次元），$\Delta\boldsymbol{m}$ は ΔMFCC（低 12 次元），E は RMS（1 次元）を表す．RMS とは音のエネルギーを表す特徴量であり，標本数 n と時刻 x の波形の振幅 z_x を用いて式（8）のように表せる．

$$E = \sqrt{\frac{1}{n}\sum_{x=1}^{n} z_x^2} \tag{8}$$

$\overline{\boldsymbol{m}}$，$\boldsymbol{\sigma}$ は半小節間での MFCC の平均と標準偏差を表す．また，${}^jm,m^j$ はそれぞれ，小節の前半区間における特徴量，小節の後半区間における特徴量を表す（**図 3**）．N は（半拍分の時間）/（分析窓幅）であり半拍区間で算出される特徴量の総数を表す，α は繋げる小節の順序が出来るだけ元の楽曲を保存するための重み係数である．そして，楽曲の第一小節の終わりから指定箇所を経由し最終小節の始まりにかけて，コストが最小となる経路を動的計画法により決定し，経路上の小節を繋げることで BGM を生成する．こ

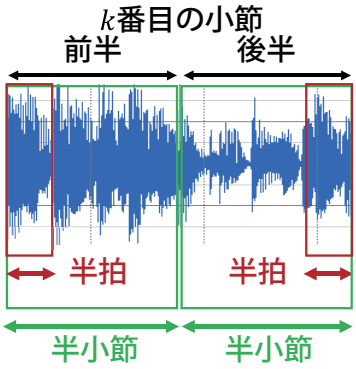

図 3　音響特徴量の抽出

こで，楽曲のサンプリング周波数は 44.1kHz，分析窓幅・シフト幅は 10ms とし，抽出した特徴量は楽曲単位で標準化してある．また，楽曲の小節構造の情報は手動で作成された既存のアノテーション情報を用いた．

4.2　ユーザの指定に基づく盛り上がりの反映

楽曲の盛り上がりを反映した BGM の生成は，4.1 節で述べた手法と同様に動的計画法を用いた楽曲小節単位での切り貼りによる補間によって行う．ただし，ユーザによる盛り上がりの指定を反映させるために小節間の繋がりやすさを表す式 (1) にユーザの指定した盛り上がり度合いと楽曲の盛り上がり度合いとの差を考慮した式 (9) を用いて評価を行う．

$$C'_{jk} = C_{jk} + |S_k - S_U| \tag{9}$$

S_k は k 番目の小節の RMS の平均を表す．S_U はユーザが指定する定数であり，大きくするほど盛り上がりが大きい小節が選ばれやすくなる．本稿では，盛り上がりを表す特徴量として，便宜的に RMS を用いているため，音響信号上での振幅の大きさが反映されている．今後はより高次な盛り上がり特徴の導入について検討する予定である．

4.3　映像と楽曲の指定箇所の同期

小節単位での楽曲の切り貼りを行うため，生成された BGM の長さは元の楽曲の小節の長さの自然数倍となる（**図 4**，Ⓐ）．このため，映像の指定箇所の開始時刻が小節の長さの自然数倍でない場合，生成した BGM において映像と楽曲の指定箇所の位置が最大で 1/2 小節分ずれてしまう（図 4，Ⓑ）．そこで，楽曲の第一小節の終わりから指

Vol.2015-MUS-106 No.10
Vol.2015-EC-35 No.10
2015/3/2

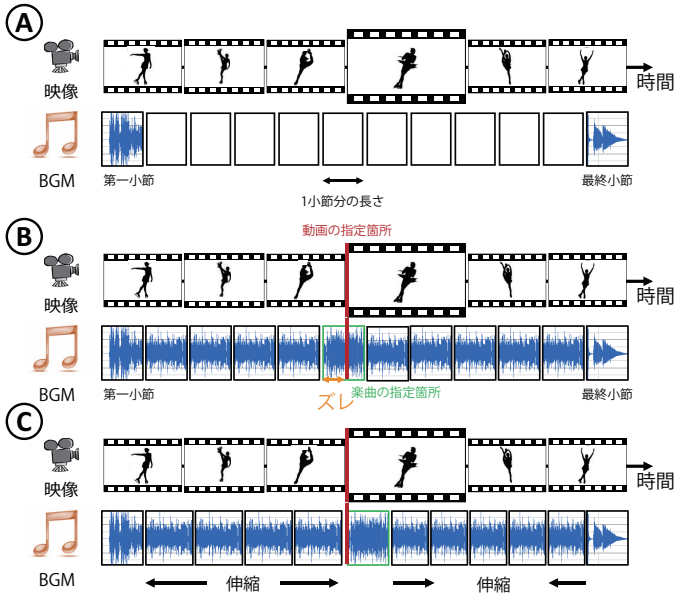

図 4　映像と楽曲の指定箇所間のズレ

定箇所の始まり，指定箇所の終わりから最終小節の始まりの区間のずれに応じて図 4，© のように生成された BGM を伸縮させる．これにより，映像と楽曲の指定箇所が同期した BGM を付加した映像の生成が可能となる．ここで，指定箇所間の長さが短い区間を伸縮させた場合，BGM のテンポが急激に変化してしまうことがある．今後，映像と BGM の両方を伸縮させることで楽曲の伸縮を減らす，テンポの変化を滑らかにする処理を加えることで急激なテンポの変化を抑えるといった改善が必要である．

5.　生成結果の評価

こだわりを反映した BGM を映像へ付加することで，映像をより印象的に演出することが可能となる．しかし，小節の切り貼りによって BGM を生成することで小節の繋ぎ目に違和感が生じてしまうと，映像を印象的に演出する効果は減少してしまう．そこで，本手法が効果的に働く楽曲の特徴を検証するための実験を行った．実験では，ポップスやジャズ，クラシックなどの様々なジャンルの楽曲に対し，本手法を用いて BGM を生成し，繋ぎ目の自然さを順位法をによって被験者に評価してもらった．

5.1　実験条件

20 代の大学生 5 人（男性 4 人，女性 1 人）の被験者に後述するデータベース [11] から制作した 10 曲の BGM を視聴してもらい，繋ぎ目の自然さに基づいて順位付けを行ってもらった（自然なほど順位は若くなる）．実験は，生成された BGM の繋ぎ目の自然さの評価が目的であること，生成された BGM を全て試聴するのは被験者の負担となることの二点を考慮し，各 BGM の長さは冒頭 30 秒とした．また，映像の指定箇所を調節し，冒頭 30 秒の間には小節の繋ぎ目が少なくとも二箇所あるようにした．

実験で用いる楽曲は RWC 研究用音楽データベース (RWC-MDB)[11] の音楽ジャンルデータベース (RWC-MDB-G-2001) の中から各ジャンル 1 曲ずつ，計 10 曲を抜粋した．楽曲の拍及び小節構造の情報は AIST Annotation for RWC Music Database (Beat structure) を用いた [12]．

表 1　順位相関係数 ρ の相関がある組の数と統計量

1%水準	5% 水準	平均	標準偏差	最大	最小
3 組	8 組	0.6448	0.1278	0.8424	0.4545

表 2　BGM に用いた楽曲と順位の中央値

曲番号・ジャンル	中央値	曲番号・ジャンル	中央値
No.72・ワールド	2	No.87・声楽	7
No.29・ジャズ	3	No.91・邦楽	7
No.38・ラテン	3	No.21・ダンス	8
No.1・ポップス	4	No.53・クラシック	9
No.8・ロック	4	No.55・行進曲	9

5.2　実験結果と考察

5.2.1　順位結果の信頼度評価

被験者の組み合わせ（10 組 $= {}_5C_2$）全てに対して順位相関係数を計算し，1%・5%水準で相関がある組の数と統計量を求めた．被験者間の評価に高い相関があれば，順位の結果には信頼性があり，順位の中央値が上位の楽曲は自然な繋がりであると言える．そこで，被験者二人がつけた順位の相関を Spearman の順位相関係数 [10] を用いて算出する．同順位のない順序ベクトル $\boldsymbol{a}$,$\boldsymbol{b}$ の順位相関係数 ρ は式 (10) で表される．

$$\rho = 1 - \frac{6}{N^3 - N} \sum_{i=1}^{N} (a_i - b_i)^2 \qquad (10)$$

N は楽曲数，a_i, b_i は順序ベクトル $\boldsymbol{a}$,$\boldsymbol{b}$ における i 番目の要素の順位である．今回の実験では楽曲数が 10 曲であるため，$\rho \geq 0.7333$，$\rho \geq 0.5636$ ならばそれぞれ 1%水準，5%水準で有意な相関がある．順位相関係数の 1 **表 1** より順位相関係数の平均が 5%水準の値を超えていることが分かる．また，その組み合わせの数も 10 組中 8 組であり，5%水準の値を超えた組み合わせの割合が高いことが分かる．この結果から，被験者間の順位の評価には高い相関があり，順位の中央値が上位である楽曲ほど小節の繋ぎ目は自然であることが分かる．

5.2.2　考察

BGM に用いた楽曲の各ジャンルとその順位の中央値を **表 2** に示す．表 2 より順位の中央値が上位である楽曲はワールド，ジャズ，ラテンであった．これらの楽曲は同じ音色の繰り返しが多く類似した小節が多く存在するため，小節の繋ぎ目が自然になりやすかったと考えられる．一方，順位の中央値が下位である楽曲はクラシック，行進曲であった．これらの楽曲は，音色やテンポの変化が多い楽曲で

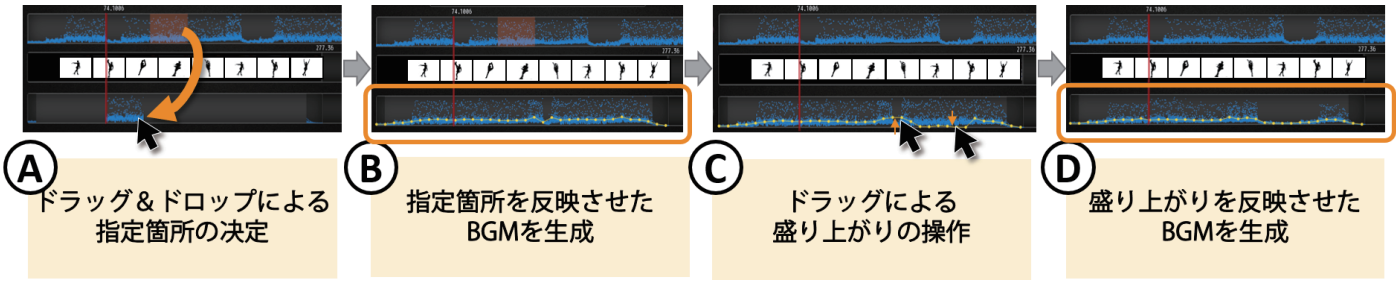

図 5　こだわりを反映させた BGM の生成操作例

あったため，小節の繋ぎ目で生じる違和感が大きくなりやすかった．これらの問題については，音楽構造やテンポを考慮したコスト関数を設計することで，対応可能となると考えている．また，現段階では歌声の有無を考慮していないため，小節の繋ぎ目で歌声が出現・消失することがある．ポップス及びロックの楽曲には歌声が存在するが，今回実験に用いた BGM 中では歌声区間と非歌声区間を繋いだ小節が現れなかったため上位に評価されたと考えられる．以上より，本手法は同じメロディが多いインストゥルメンタルの楽曲を利用した場合，特に有効であると考えられる．

6. BGM 付加支援インタフェース

本手法では，映像と楽曲の始端と終端を一致させながら，映像と楽曲の指定箇所を同期させた BGM を生成する．しかし，生成された BGM は必ずしもユーザ好みの楽曲の盛り上がりが反映された BGM になるとは限らない．そこで一度生成された BGM に対して，ユーザが楽曲に盛り上がりを指定することでユーザの好みを反映した BGM を生成できるようにするインタフェースを提案する．これにより，ユーザは直感的な操作で，指定箇所を同期させながら楽曲の盛り上がりを反映させた BGM を生成し映像へ付加することができるようにする．

6.1 楽曲編集機能

本システムのインタフェースの画面を図 **6** に示す．基本的な機能として，再生中の映像の表示（図 6，Ⓐ），楽曲や映像の読み込み，BGM を付加した動画の書き出し（図 6，Ⓑ），映像の再生や停止（図 6，Ⓒ），入力楽曲の RMS の表示（図 6，Ⓓ），入力映像のサムネイル画像の表示（図 6，Ⓔ），生成された BGM の表示（図 6，Ⓕ）がある．BGM を生成する手順として，初めにユーザは入力楽曲の利用したい箇所とそれを付加する映像の箇所をドラッグとドロップによるマウス操作によって指定する．ユーザの指定に基づき指定箇所以外の区間が自動補間された BGM が生成される（図 6，ⒶⒷ）．さらに，生成された BGM がユーザの好みに合わない場合，ユーザは黄色の線（図 6，Ⓖ）で描画されている BGM の盛り上がりをドラッグによって調節することで，ユーザの盛り上がりの好みを反映した BGM を生成することができる（図 **5**，ⒸⒹ）．

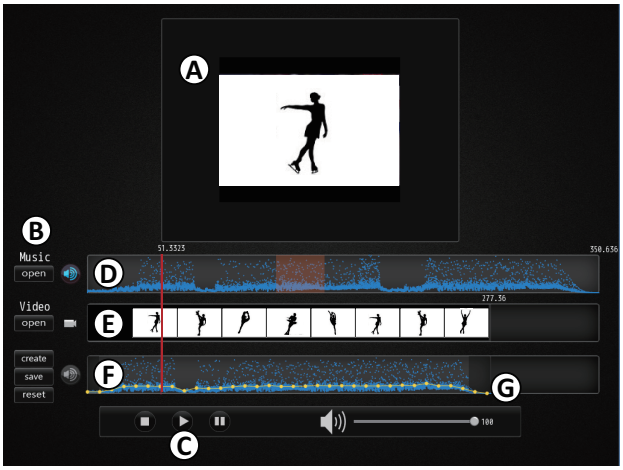

図 6　インタフェース画面

7. まとめと今後の課題

本稿では，映像と楽曲の始端と終端を一致させながら，映像と楽曲の指定箇所が同期した BGM を映像へ付加させる手法を提案した．こだわりを反映させた BGM を映像へ付加するには技術や手間暇が必要である．しかし，本システムを用いることで，楽曲の編集経験が無いユーザでも容易にこだわりを反映させた BGM を制作し映像へ付加可能となった．しかし，歌声がある楽曲や，楽曲中でテンポや音色の変化が大きい楽曲に本手法を適用した場合，生成した BGM の繋ぎ目に違和感が生じてしまう問題があった．今後，テンポの変動や歌声がある楽曲に本手法を適用可能にするようなコスト関数の設計を行いたい．さらに，生成される BGM の自然さの向上に取り組みたい．また，生成された BGM に対してユーザが盛り上がりの流れを指定することで，楽曲の盛り上がりを反映させ BGM を再編集するシステムを提案した．しかし，今回は簡易的に盛り上がりの算出を行ったので，今後は盛り上がりの算出で用いる手法や音響特徴量の検討を行いたい．インタフェースによる直感的な操作でこだわりや盛り上がりを反映させた BGM を生成し映像への付加を支援するシステムを目指した．生成させる BGM の評価は今後の課題であり，既存ソフトによる編集結果と比較評価することで本システムの性能を評価し

たい. 今後，音楽と映像の同期尺度 [13] を導入することで，人が知覚的に映像と同期していると感じるような BGM の付加に取り組みたい. これにより，ユーザのこだわりや盛り上がりを反映しながら，映像にシンクロした BGM の付加を支援するシステムの実現を目指したい.

謝辞 本研究の一部は，JST CREST「OngaCREST プロジェクト」の支援を受けた.
また，RWC 研究用音楽データベース [11] を使用した.

参考文献

[1] 岩宮眞一郎：オーディオ・ビジュアル・メディアを通しての情報伝達における視覚と聴覚の相互作用に及ぼす音と映像の調和の影響，音響学会誌，Vol.48，No.9，pp.31–39 (1992).
[2] Yoon, J. C., Lee, I. K. and Byun, S.:Automated music video generation using multi-level feature-based segmentation, Handbook of Multimedia for Digital Entertainment and Arts, pp.385–401 (2009).
[3] 小野佑大，石先広海，帆足啓一郎，小野智弘，甲藤二郎：音楽のムード分類結果を利用したホームビデオへの自動 BGM 付与・同期手法，情報科学フォーラム講演論文集，Vol.9，No.2，pp.295–296 (2010).
[4] Jiashi, F., Bingbing, N. and Shuicheng, Y.:Auto-generation of professional background music for home-made videos, Proceedings of the 2nd International Conference on Internet Multimedia Computing and Service, pp.15–18 (2010).
[5] Chu, W. T., Tsai, and S.Y.:Rhythm of motion extraction and rhythm-based cross-media alignment for dance videos, IEEE Transactions on Multimedia, pp.129–141 (2012)
[6] Zhang, W., Xing, L. and Huang, Q.:A System for Automatic Generation of Music Sports-Video, IEEE International Conference on Multimedia and Expo, pp.1286–1289 (2005)
[7] Wang, J., Chng, E., Xu, C., Lu, H. and Tian, Q.:Generation of personalized music sports video using multimodal cues, IEEE Transactions on Multimedia, pp.576–588 (2007)
[8] Lu, L. and Zhang, H, J.:Automated extraction of music snippets, Proceedings of the 7th International Conference on Music Information Retrieval, Proceedings of the eleventh ACM international conference on Multimedia, pp.140–147 (2003).
[9] 白鳥貴亮，中澤篤志，池内克史：音楽特徴を考慮した舞踊動作の自動生成，電子情報通信学会論文誌 D，Vol.90，No.8，pp.2242–2252 (2007).
[10] Kendall, M. and Gibbons, J. D.:Rank Correlation methods, 5th edition, p.260, Oxford University (2006).
[11] 後藤真孝，橋口博樹，西村拓一，岡隆一：RWC 研究用音楽データベース：研究目的で利用可能な著作権処理済み楽曲・楽器音データベース，情報処理学会論文誌，Vol.45，No.3，pp.728–738 (2004).
[12] Masataka Goto.:AIST Annotation for the RWC Music Database, Proceedings of the 7th International Conference on Music Information Retrieval, pp.359–360 (2006).
[13] 平井辰典，大矢隼士，森島繁生：既存音楽動画の再利用による音楽に合った動画の自動生成システム，情報処理学会論文誌，Vol.54，No.4，pp.1254–1262 (2013).