Vol.2012-MUS-95 No.10
2012/6/3

隠れマルコフモデルに基づく 既存コンテンツ学習による音楽動画自動生成システムの提案

大矢 隼士[†1,a)]　森島 繁生[†1]

概要：インターネットの動画共有サイト上に存在するアマチュア制作の音楽動画を再利用することにより，自動的に音楽動画を生成するシステムを提案する．この音楽動画は，既存の音楽にゲームやアニメなどの映像を切り貼りして制作されたものであり，MAD 動画と呼ばれている．本稿では，以前筆者らグループが提案した DanceReProducer の学習手法を，マルコフ連鎖を使うことにより映像の時系列情報を考慮できるように改善し，Forward Viterbi アルゴリズムを用いて動画生成をおこなう．提案システムは，まずインターネット上にアップロードされている MAD 動画を大量に取得し，データベースとする．その後，データベースの動画から音楽特徴量，映像特徴量を抽出し一小節ごとにまとめ，楽曲の構造情報やテンポの推定をおこなう．次に，各特徴量をクラスタリングし，状態変数を音楽特徴量，潜在変数を映像特徴量として，潜在変数のマルコフ連鎖モデルを使用して学習する．動画の生成は，任意の楽曲（入力楽曲）に対し，学習した同調関係から最も入力楽曲と同調する映像をデータベースから選び出し，切り貼りすることで新しい動画を自動的に生成している．

キーワード：動画生成，マルコフモデル，Forward Viterbi アルゴリズム

Automatic Music Video Generating System by Learning Existing Contents Based on Hidden Markov Model

Ohya Hayato[†1,a)]　Shigeo Morishima[†1]

Abstract: We propose automatic music video creating system by reusing music videos that are created by amateur users and existed on the video sharing service. These music videos are created by combining existing music with frames of video games, anime, and so on and are called "MAD" movie. In this paper, we improved learning method of DanceReProducer that is the system our group proposed previously, by considering time series information of videos by using Markov chain, and we automatically generated MAD movie by using Forward Viterbi algorithm. First, we get a lot of MAD movie from the web and makes database. Next, music feature and video feature are extracted from MAD movies in the database, and gathered per bar. Then, each feature is clustered and learned by Markov model of latent variable as state variable is music feature and latent variable is video feature. At last, movie is automatically generated by selecting video frame, which is most synchronized with input music in the database, and combining input music with video frames.

Keywords: video generation, Markov model, Forward Viterbi algorithm

[†1] 現在，早稲田大学/CREST
Presently with Waseda Uniersity
[a)] hayato-o@ruri.waseda.jp

1. はじめに

近年，MAD 動画とよばれる，ゲーム画面やアニメ動画など，既存の動画素材から音楽に合う映像を探しだし，映像を切り貼りすることで制作された動画が，インターネット上で人気を博している．MAD 動画の多くは，アマチュアによって制作されおり，動画共有サイト上にアップロードされ人々に楽しまれている．特に制作された映像が音楽と高く同調している MAD 動画は，視聴者から高い評価を受け多く再生されている．そのため制作者も，制作した動画の「再生数」を一つの評価のバロメータと捉えて制作のモチベーションにしている．しかし，音楽と映像の同調関係についての知見は解明されていない点が多く，多くの視聴者から評価される動画を制作したいと考えても，難しい課題であった．

また，こうした創作文化の展望から，コンテンツを再利用した創作が容易にできる環境に対する要請が高まっており，音楽では創作支援サービスが開始されているが [6][7]，MAD 動画の創作環境を考えると，高機能な動画編集ソフトは在るものの一から手作業で制作するしかなく，手間や労力を考えると満足な環境とは言えない．

以前筆者らグループが提案した DanceReProducer[1] は，動画共有サイト上の既存のダンス動画を再利用することにより，自動的にダンス動画を生成するシステムである．しかしこのシステムは，1 小節分の音楽特徴量と映像特徴量の間の関係を学習するのみで，映像特徴量の時間的遷移を学習することはできなかった．MAD 動画は音楽作品であるとともに，映像作品でもあるため，映像の時系列情報を学習するいうことは，非常に重要なファクターだと考えられる．

そこで本研究では，潜在変数のマルコフ連鎖と Forward Viterbi アルゴリズムを用いて映像の時系列情報を学習することで DanceReProducer を改善し，主観評価実験により本研究と DanceReProducer，人手で制作された動画それぞれの音楽と映像のシンクロ度合いを比較した．

2. 関連研究

DanceReProducer では既存の動画コンテンツを取得し，それらの音楽と映像の同調関係を学習する．その後，新たな楽曲（以下入力楽曲）に対して，学習した音楽と映像の対応関係をもとに，音楽にあった映像を選択し，つなぎ合わせることで MAD 動画を自動生成した．音楽と映像の同調関係の学習には，アマチュアにより制作された MAD 動画を教師データとして用い，短時間で特徴量を抽出した後，楽曲の小節単位でまとめておこなっている．この際，映像特徴量と音楽特徴量の学習は一対一対応でされているため，映像の時系列情報を考慮できていない．

映像編集のために時系列情報を学習した研究として，姜らの研究 [2] がある．姜らは，専門家により制作された編集映像とその素材映画から学習した編集技術に基づいて，映画の自動映像編集をおこなった．編集映像を構成するショットを，ショット長，動き，輝度といった画像特徴量の量子化に基づきシンボル化し，その時系列パターンを HMM により学習した．ただし，この研究では映像に付加している音に関しての学習はされていない．

そこで本研究では，潜在変数を映像特徴量，観測変数を音楽特徴量としたマルコフ連鎖で学習し，Forward Viterbi アルゴリズムを用いて動画生成をおこなう手法を提案する．

3. 研究概要

本研究の処理フローを図 1 に示す．本研究は，「データベース構築フェーズ」，「学習フェーズ」，「動画生成フェーズ」の 3 つから構成される．

データベース構築フェーズでは，既存の動画群から，音楽および映像特徴量を抽出し，楽曲の構造・テンポ・小節線を推定する．その後，1 小節ごとに各特徴量をまとめる．学習フェーズでは，各特徴量のクラスタリングをおこなった後，マルコフ連鎖により，HMM パラメータを学習する．本研究で使用した特徴量を表 1 に記載する．本研究では，DanceReProducer と同様の特徴量を使用している．

動画生成フェーズでは，まず映像を付与したい音楽（以下入力楽曲）に対して，音楽特徴量の抽出，テンポ・楽曲構造・小節線の推定をおこない，1 小節ごとに特徴量をまとめる．その後，学習フェーズでもとめた HMM のパラメータを使い，Forward Viterbi アルゴリズムにより映像系列を決定する．1 小節の映像特徴量を推定するために，その小節の音楽特徴量に加え，1 小節前の映像特徴量の状態を鑑みることで，映像特徴量の時間的遷移を考慮した動画が生成できる．

表 1 使用した特徴量

	次元	音楽特徴量		次元	映像特徴量
アクセントに関する特徴量	1.-4.	サブバンド毎のパワー		1.	オプティカル・フローの微分値
	5.	Spectral Flux		2.	輝度値ヒストグラムの微分値
印象に関する特徴量	6.	Zero-crossing rate		3. 4.	色相の平均と分散
	7.-19.	MFCC		5. 6.	彩度の平均と分散
				7. 8.	明度の平均と分散
				9.-20.	Discrete Cosine Transform

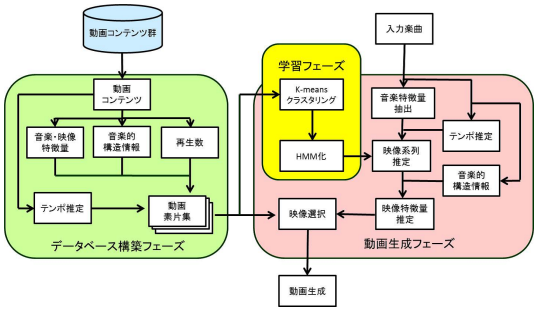

図 1 システム全体のフローチャート

4. データベースの構築

まず，データベースの構築について述べる．システムにデータベース化する動画コンテンツ群を与え，前処理として，各動画の映像のフレームレートを 30fps に，音楽のサンプリングレートを 44.1kHz にリサンプリングする．続いて，映像の 1 フレームごとに音楽と映像の特徴量を抽出していく．また，動画に付加されている楽曲の小節線の位置を推定することで，取得した各特徴量を小節単位にまとめいく．

4.1 音楽特徴量の抽出

西山らの研究 [4] と楽曲ジャンル分類に関する先行研究 [5] を参考に，アクセントおよび印象に関する特徴量を決定した．本研究で使用した音楽特徴量は表 1 の通りである．アクセントに関する特徴量としては，主に楽曲のパワーとその時間変化を表現するために，フィルタバンクごとのパワー（4 次元）と Spectral Flux（1 次元）を用いている．本研究では DanceReProducer にならい，フィルタバンクのバンク数を 4 とした．印象に関する特徴量としては，楽曲の音色に関連した Zero-crossing rate（1 次元）と MFCC（Mel-Frequency Cepstral Coefficients）の直流成分と低次 12 項を用いる（13 次元）．これにより，音楽のアクセントと印象に関する計 19 次元のフレーム特徴量を抽出する．

なお，分析窓のシフト幅は映像のフレーム特徴量と対応を取るため，1470 点（= 44100 Hz ／ 30 fps）とし，窓長はシフト幅よりも長くするため，2048 点としている．

4.2 映像特徴量の抽出

映像特徴量は，西山らの研究 [4] を参考に，アクセントおよび印象に関する特徴量を使用している．本研究で使用した音楽特徴量は表 1 の通りである．アクセントに関する特徴量としては，画面の動きやダンス動作とそれらの時間変化や画面の切り替わりを表現するために，オプティカル・フローと輝度値の時間微分の平均値を用いている（各 1 次元）．オプティカル・フローはブロックマッチング法を用い，ブロック数 64 × 48，シフト幅 1，最大シフト幅を 4 として計算している．印象に関する特徴量としては，映像の雰囲気を表現するために，全画素における色相，彩度，明度，それぞれの値の平均と標準偏差を用いる（全 6 次元）．また，映像全体の印象を表現するため，二次元の離散コサイン変換（DCT: Discrete CosineTransform）の係数 12 次元（＝画面横軸方向の低次 4 項×画面縦軸方向の低次 3 項）を用いている．これにより，映像のアクセントと印象に関する計 20 次元のフレーム特徴量を抽出する．

なお，前処理として，画面サイズを 128 × 96 にリサンプリングをおこなっており，分析窓のシフト幅は，映像のフレームレートである 1 フレームごと（1/30 s）としている．

4.3 小節線の推定

小節線の推定には，音響信号のパワーの相関函数に基づく方法で計算をおこなっている．まず，入力音響信号のパワーの自己相関函数のピーク時刻を求める．これはパワーの周期性を表すため，これを 1 拍の時間長としてテンポ推定する．ただし，倍テンポ誤りや半テンポ誤りを回避するため，テンポに 90〜180bpm の制限を設けて推定している．テンポ推定の例を図 2 に示す．

続いて，推定されたテンポから 1 拍ごとの時刻にピークを持つパルス列を生成し，それと入力音響信号のパワーの相互相関函数を計算してピーク時刻を求める．これは，楽曲中の 1 拍目の時刻を表すとする．これらの結果から，1 拍目を小節線の開始位置とみなし，また 4/4 拍子を仮定して，機械的に小節線の位置を決定する．

4.4 小節ごとの特徴量のまとめ

時系列情報を考慮した小節ごとの特徴量のまとめ方として，DCT を用いている．フレーム特徴を 16 点でリサンプリングし，この 16 点に対して DCT をかけ，そのうちの低 4 次元を使用する．DCT の低 4 次元を逆変換したものを図 3 に示す．この結果より，は後半にかけての特徴量の上昇が再現されていることが確認できる．なお，リサンプリングによる時系列変化の損失については，BPM が 60 で 30 サンプル程度であるため，リサンプルすることでの損失は少ないと考えられる．これにより，音楽特徴量は 76 次元，映像特徴量は 80 次元となる．

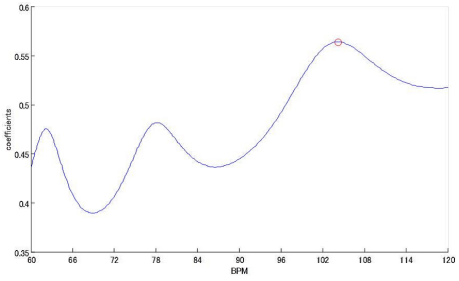

図 2 パワーの自己相関関数による BPM の推定

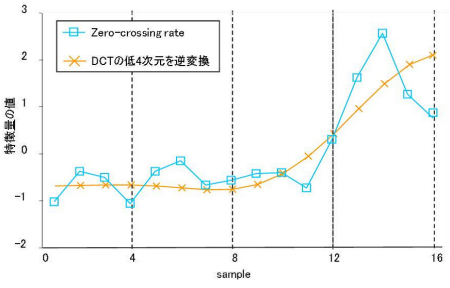

図 3 Zero-crossing rate と DCT の低 4 次元の逆変換結果

5. 音楽と映像の同調関係の学習

音楽と映像の同調関係の学習は，インデクス素片集に格納された音楽特徴量と映像特徴量を用いて HMM によりおこなう．まず，前処理としてそれぞれの特徴量を主成分分析し，特徴量の直交化，及び次元削減をおこなう（累積寄与率 95 % で削減）．これにより，音楽と映像の小節特徴量はそれぞれ，76 次元と 80 次元であったものが，62 次元と 64 次元へ削減される．

5.1 音楽特徴量と映像特徴量のクラスタリング

HMM に基づく動画生成をおこなうため，音楽特徴量と映像特徴量をそれぞれ観測変数と潜在変数の状態に見立てる必要がある．そこで，音楽特徴量と映像特徴量それぞれにクラスタリングをおこない，各クラスタを観測変数と潜在変数の状態とした．クラスタリングには k-means クラスタリングを用いた．

本稿では，クラスタ数は 10 としている．クラスタ数を増やすほど最適な映像を選び出すことが可能になると考えられるが，クラスタ数を増やしすぎると Forward Viterbi アルゴリズムを用いて最適な映像時系列を求める際にアンダーフローを起こしてしまう．本研究で扱う動画・生成する動画の長さは 3 分から 5 分程度，クラスタ数（= 楽曲の小節数) は 90 小節から 150 小節程度のものである．

クラスタリングの後，説明変数を映像特徴量、目的変数を音楽特徴量としてクラスタごとに線形重回帰をおこない，回帰係数 A を求める．この回帰係数 A は動画生成時に使用するもので，後ほど述べることにする．

5.2 HMM に基づく音楽と映像の同調モデル

図 4 にマルコフモデルによる音楽と映像の学習モデルを示す．ミュージックビデオにおける音楽特徴量のクラスタを観測データ O，映像特徴量のクラスタを潜在変数 x の状態として，潜在変数のマルコフ連鎖モデルに当てはめる．このマルコフモデルから，潜在変数 x_i から x_j への状態遷移確率 T_{ij}，潜在変数 x のときに出力 O を出力確率 $P(O|x)$，潜在変数 x の初期状態確率 S_p を求め、HMM パラメータとする．

5.3 再生数を利用した重み付け学習

本研究の学習データは，インターネットの動画共有サイト上に存在するアマチュアによって制作された作品であるため，その品質にはばらつきが生じ，すべてを同等の学習サンプルとして扱うことは不適切であると考えられる．そこで，「再生数」を統計的な評価指標として考え，それによる重み付け学習をおこなう．

ここで s を再生数とした場合の重み ω を式 (1) により定義した．

$$\omega = \alpha \cdot \lfloor \log(s) \rfloor + \beta \tag{1}$$

再生数が 1,000 再生程度で重みが 1，10,000 再生程度で重みが倍に，100,000 再生程度でその 3 倍の重みとなるように，$\alpha = 1.0$，$\beta = -2.0$ に設定している．また，再生数が 1,000 再生未満のものは学習の対象外とする．

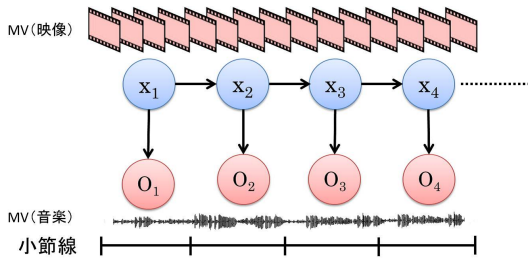

図 4 音楽と映像の学習モデル

6. 動画の生成

任意の入力楽曲を入力とした音楽動画生成について述べる．まず，前処理として，データベース構築フェーズと同様に，入力楽曲の音楽特徴量の抽出，楽曲構造とサビ箇所の取得，小節線の推定をおこなう．

続いて，5 章で述べたことと同様に，音楽特徴量を小節ごとにクラスタリングをおこなう．その後，Forward Viterbi アルゴリズムにより，映像特徴量のクラスタを推定する．この概略を図 5 に示す．

最後に，映像特徴量クラスタと入力楽曲の楽曲構造，音楽特徴量をもとに映像特徴量を求め動画を生成する．この概略を図 6 に示す．

6.1 映像特徴量クラスタの推定

映像特徴量クラスタの推定には，Forward Viterbi アルゴリズムを用いる．この推定には，5 章で求めた HMM パラメータ (初期状態確率 S_p，状態遷移確率 T) と状態 Y(クラスタ 1〜10)，観測データ O を使って，隠れ状態系列 $x_0, x_2, \cdots, x_N$ は，以下の漸化式 (2)，(3) で表される．

$$V_{0,k} = P(O_0|k) \cdot S_p \tag{2}$$

$$V_{n,k} = P(O_n|k) \cdot \max_{x \in X}(T_{y,k} \cdot V_{n-1,y}) \tag{3}$$

ここで，$V_{n,k}$ は最尤状態系列である．

6.2 映像特徴量クラスタを用いた動画生成

この最尤状態系列 $V_{n,k}$ と入力楽曲の楽曲構造情報を用いて，生成動画のシーン構造を決定する．シーンの切り替えは以下の条件でおこなう．

条件 1： x_m と x_{m-1} が異なるクラスタだった場合
条件 2： x_m と x_{m-1} の間で楽曲構造が変わったとき

上記条件をもとにシーンを決定した後，5 章で算出した回帰係数 A と説明変数である音楽特徴量から，映像特徴量を求める．求めた映像特徴量とデータベースの映像特徴量の距離をマハラノビス汎距離で計り，最も距離が近い特徴量を持つシーンを映像として貼り付ける．

映像生成の際、不自然に速い映像や遅い映像が生成されることを防ぐために，テンポが入力楽曲と ±20 % 以上異なるデータベースの映像は選択候補から除外し，また，一度使われた映像は，同じ動画で二度と使われないようにしている．

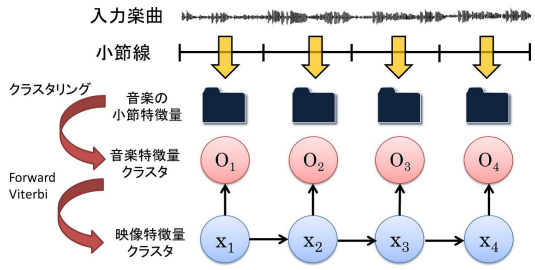

図 5 Forward Viterbi アルゴリズムを用いた映像クラスタ推定

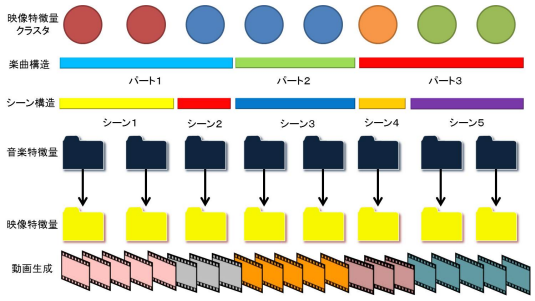

図 6 映像クラスタから動画の生成

7. 生成結果と考察

本研究を用いて動画を生成した結果，音楽の印象にマッチした映像が付加された．ダンス音楽を入力楽曲とした場合，ネオンが煌びやかな映像やライトの光が印象的な映像が付加された．また，ポピュラー音楽を入力楽曲とした場合，楽曲の構造が切り替わると音楽の印象が大きく変化するが，楽曲構造ごとに音楽の印象にマッチした映像が付加された．特に，楽曲の構造が切り替わり，曲が盛り上がる部分では，輝度が高く派手な印象の映像が見受けられた．ボーカルが入っていないジャズ音楽を入力としてみると，暗めな映像と明るめな映像の両方が付加された．この明るめな映像は，彩度があまり高くないもので，ジャズ音楽の落ち着いた雰囲気と合っている印象を受けた．

DanceReProducer で同様の楽曲を用いて動画を生成したところ，曲調の変化が少ないダンス音楽では，彩度の低い映像が付加された．また，ポピュラー音楽では，比較的明るい映像が付加されているが，楽曲構造におけるメリハリがあまり感じられなかった．ジャズ音楽では，暗めの映像を付加しているものの，同一の動画からばかりシーンを選択している．これにより，時系列情報を学習した動画生成をおこなうことで，より音楽の印象と合った映像が付加されやすくなると考えられる．

8. 主観評価実験

8.1 実験内容

入力楽曲 4 曲に対して，本研究と DanceReProducer を使用して動画計 8 個を生成し，同じ曲を使用して人手で制作された動画 4 個を加え，計 12 個の動画の比較をおこなった．比較する動画は，DanceReProducer との比較という観点から，全てダンス動画としている．動画の生成には，90 個の既存動画をデータベースとして，学習と動画生成に使用した．人手で制作された動画は，再生数 5 万回以上（完成度が高いと思われる）のものを選択した．5 万回というのは，データベース中の動画の中でも高い数値である．DanceReProducer および本研究では，動画の完成度の尺度として再生数を用いているため，完成度の高い動画と比較することを意味している．

20 代男性 15 名にこれらの動画を視聴してもらい，音楽と映像のシンクロ度合いを評価してもらった．評価データは，動画の時間情報と評価を記録したものである．

8.2 実験結果

評価結果を図 7 に提示する．”User-generated”は人手で制作された動画，”DanceReProducer”は DanceReProducer，”Our method”は本研究を示している．図 7 より，人手で制作された動画と自動生成したものでは「かなりシンクロしている」という割合が大きく異なる．これは，小節線推定誤りなど特徴量の検出において誤差が生じたため，ダンスと映像がずれてしまったためだと考えられる．DanceReProducer と本研究を比較すると，本研究では DanceReProducer に比べ，「かなりシンクロしている」と「シンクロしている」の合計の割合が改善された．7 章で述べた「音楽と映像の印象が合っている」と感じることが，ここに反映されていると考えられる．

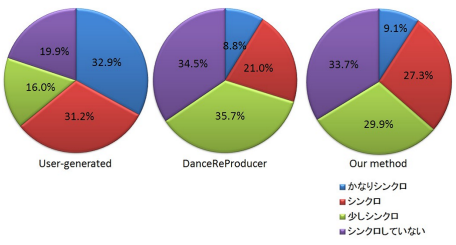

図 7 全動画時間に対する各評価時間の割合

9. むすび

本研究では，MAD 動画における音楽と映像の同調関係をマルコフ連鎖により学習し，Forward Viterbi アルゴリズムを用いて，新たな MAD 動画を自動生成した．評価実験の結果，動画生成としての精度はあがったものの，今回の評価方法では時系列を学習したことによる明示的なデータを示すことはできない．したがって，今後の課題として評価手法の検討が必要であり，「シンクロ」という設問方法や「音楽と映像の印象が合っている」という客観評価尺度の検討をおこなっていきたい．

また，本研究では，DanceReProducer では実装されていた GUI は実装されていない．自動生成した動画の気に入らない部分をユーザが気軽に編集できる GUI を作成し，その編集記録を学習することで，最終的には一度の試行で完成度の高い映像を生成出来るようなシステムへ改善できる可能性がある．これによって得られる学習データは，音楽特徴量と映像特徴量が物理的に最も整合する「最適解」ではなく，人間が定性的に「音楽と映像がこの程度あっていれば良い」と感じる指標となる「満足解」であると考えられる．自動動画生成という観点でいえば，動画全体で音楽と映像の最適なパラメータを合わせるのではなく，動画の部分部分で人間が満足できるものをつなぎ合わせることで，人間が動画を視聴した際に自然に感じる動画の自動生成が可能になると考えている．そのため，今後は満足解を発見する GUI と自動動画生成システムの改善を検討してきたい．

参考文献

[1] Tomoyasu Nakano, Sora Murofushi, Masataka Goto, and Shigeo Morishima, “ DanceReProducer: An Automatic Mashup Music Video Generation System by Reusing Dance Video Clips on the Web ” Proc. SMC2011, pp.183-189, 2011.
[2] 姜国臻，新田直子，馬場口登， ” 映像編集のための事例映像に基づく素材映像からのショット列生成”，電子情報通信学会技術研究報告， PRMU2007-265, pp.127-132, 2008.
[3] M. Goto, “A Chorus-Section Detection Method for Musical Audio Signals and Its Application to a Music”, IEEE Transactions on Audio, Speech, and Language Processing，pp.1784-1794，2006.
[4] 西山正紘，北原鉄朗，駒谷和範，尾形哲也，奥乃博，“マルチメディアコンテンツにおける音楽と映像の調和度計算モデル”，情報処理学会研究報，2007-MUS-069，pp.111-118，2007.
[5] G. Tzanetakis，P. Cook，“Musical Genre Classication of Audio Signals”, IEEE Transactions on Speech and Audio Processing，pp.293-302，2002
[6] ソニーコンピュータサイエンス研究所, Music Mashroom, http://www.musicmashroom.com/
[7] The Echo Nest Corporation，the echonest, http://echonest.com/