ダンスモーションにシンクロした 音楽印象推定手法の提案とダンサーの表情自動合成への応用

朝比奈わかな[†1,a)]　岡田成美[†1]　岩本尚也[†1]　増田太郎[†1]　福里司[†1]　森島繁生[†2]

概要：近年，3DCG 制作ツール（MikuMikuDance 等）の普及により，楽曲に合わせて CG キャラクタを踊らせる動画作品が増加傾向にある．このようなダンス動画においてキャラクタの表情は作品全体の印象に大きく影響する．例えば，楽曲やダンスモーションの印象と全く異なる印象の表情を付与した場合，その動画は違和感のある作品となってしまう．また，楽曲の印象が一定でも，ダンスモーションの激しさの度合いやキャラクタの姿勢などによって，印象が大きく左右される場合がある．そのため，作品の印象を決定する要素として，楽曲の印象だけでなく，ダンスモーションの情報も考慮する必要がある．そこで我々は，ダンス動画の印象についての主観評価実験に基づき，音響特徴量とモーション特徴量を用いて重回帰分析を行うことで，ダンスモーションにシンクロした楽曲印象推定を可能にした．

1. はじめに

近年，急速な科学技術の発展により，DJ（Disc Jockey）や VJ （Video Jockey）などをはじめとする多くの人が，音楽を聴くだけでなく，自ら音楽を再編集したり，音楽に映像を付与したりすることで音楽を新たな側面から楽しむ文化が形成されつつある．音楽を映像化し，聴覚の情報に視覚からの情報を加えることで，人々は音楽の印象をより明確に感じ取ることができるようになる．このような背景から，最近では，多くのメディアプレイヤーに再生中の音楽に合わせて様々な色の球や線などが動くような幾何学的なビジュアル化システムが用いられるようになった．しかし，本来球や線自体は具体的な意味を持たないため，楽曲に含まれる感情や印象を表現するのは困難である．一方で，楽曲に合わせて CG キャラクタを踊らせるダンス動画は，キャラクタという擬人化表現を用いることで，楽曲に含まれる感情をより直観的に伝えることができる．擬人化表現の例として，3DCG 制作ツール（MikuMikuDance 等）がアマチュアの制作者にも無償配布されるようになったことから，ダンス動画作品数は年々増加傾向にある．さらに，動画共有サイトなどでお互いの作品を見せ合うなど，その楽しみ方は多様化している．しかし，3DCG 制作ツールは，一つの作品を制作する際に 1 フレーム毎にポーズを手動で決定する必要があるなど，未だ多くの労力や時間を要する．さらに，動画作成時の手間から，ダンスキャラクタの表情を的確に表現している作品は少なく，その不自然さから見応えに欠けるという問題があげられる．表情は，一般的に感情や情緒を表す非言語コミュニケーションの一つとされており，相手に感情を円滑に伝える重要な役割を果たす．そのため，ダンス動画においてもキャラクタに自然な表情を付与することで，楽曲に含まれる感情や印象をより明確に表現することができる．しかし，表情を付与するために，楽曲やダンスに関する専門的知識が必要であり，それらを取得することは容易ではない．このことから我々は，楽曲やダンスの専門的知識を持たない人でも少ない労力でダンスキャラクタの自然な表情付けができるシステム構築が必要であると考えた．

ダンスキャラクタの表情を決定する要素として，まず楽曲の印象が考えられる．楽曲は感情と深い結びつきがあるため，ダンス動画の楽曲の印象を推定することで，作品全体の印象を把握することができる．しかし，楽曲の印象のみから表情付けを行うと，ダンスモーションに対するシンクロ感が得られない．そこで，ダンスモーションの激しさの度合いやキャラクタの姿勢といったモーションの情報を加味すれば，ダンスモーションにマッチした表情付けが可能になると考えられる．また，その他の要素として，使用するダンスキャラクタやステージなどといった要素もあげられるが，本研究では考慮しないものとする．

†1 早稲田大学/JST CREST
Waseda University/JST CREST
†2 早稲田大学理工学術院総合研究所/JST CREST
Waseda Research Institute for Science and Engineering/JST-CREST
a) 2236-wakana@fuji.waseda.jp

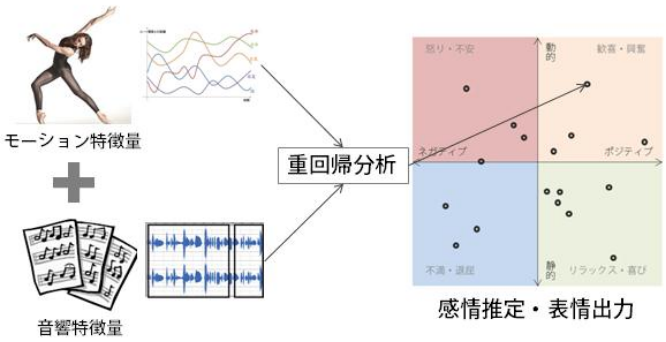

図 1. システム概要

これより，我々は，キャラクタの表情を自動合成するための，ダンスモーションに同期した楽曲印象推定方法を提案する．図 1 に本研究のシステム概要を示す．具体的に，楽曲に関しては，音色やリズムに関する特徴量を，ダンスに関しては，ラバンの身体理論に基づいたモーション特徴量を用い，それらをダンスキャラクタの表情を決定する要素とする （図 1 左側）．また，印象を表現する方法として VA 平面と呼ばれる感情平面（図 1 右側）を用い，平面上で視覚的に印象の軌跡を表現する．本手法を用いて，キャラクタに精細な表情を付与することで，よりクオリティの高いダンス動画を簡単に作ることができるようになる．

2. 関連研究

本章では，本研究を進めるにあたり，参考にした研究や比較すべき研究について説明する．

2.1 音楽のビジュアル化

多くのメディアプレイヤーのビジュアル化システムで用いられる，様々な色の球や線による幾何学的なビジュアル化表現は，球や線自体が具体的な意味を持たないため，楽曲に含まれる感情や印象を表現しきれないといった問題があげられる．そこで，Thomas ら [1] は具体的な意味を持つ 3D シェイプを用いることで音楽のビジュアル化を行った（図 2）．具体的には，音の周波数に合わせて 3D シェイプを伸縮させることで楽曲を可視化しており，また GUI 上でパラメタを調節することで 3D シェイプの変形具合を調節することもできる．しかし，3D シェイプを伸縮させること自体は具体的な感情と結びつかないため，音楽に含まれる感情要素を表現することができない．そこで，DiPaola ら [2] は，感情や情緒と深い結びつきがある表情に注目した．具体的には，楽曲の印象推定を行うことで，楽曲印象にマッチした CG キャラクタの表情アニメーションを生成している．また，表情の出力としてリアルな CG キャラクタを使用しているため，楽曲に含まれる微妙な感情変化をより明確に理解することができる（図 3）．

図 2. 出力結果例 [1]

図 3. 出力結果例 [2]

2.2 楽曲印象推定

DiPaola ら [2] は，出力する表情を決定する際に楽曲の印象推定を行っている．具体的には，Thayer モデルを基に，音量・リズム・音色に関する音響特徴量を用いて大きく 4 つの印象にクラスリングし，さらにその他の MIDI データから得られる特徴量によって，Russell のモデル [3] に基づいたより細かな印象へ分類している．この研究では，Thayer モデルや Russell の印象語などを用いることで，心理学に基づいた印象推定が可能である．しかし，この手法では MIDI データを利用した楽曲印象推定を行っているため，より普及している MP3 形式などの楽曲データに対応することができない．さらに，出力する表情の数が使用する印象語の数に限られるため，ある 2 つの印象の中間にあるような微妙な印象の違いなどは表現できない．

MIDI データを利用せず，あらゆる感情を表現することのできる VA 空間を用いて印象抽出を行っている研究として中西らの研究 [4] があげられる．中西らは，MP3 形式などの楽曲の音響特徴と歌詞各々から印象推定を行い，それらの印象の軌跡を VA 平面上にプロットすることで，歌詞と音響特徴の印象の類似度合いや楽曲印象の時間的構造を表現した．具体的には，音響特徴には重回帰分析，楽曲の歌詞には PLSA を用いることで印象推定を行った．しかし，この研究では歌詞の付いていない楽曲には対応することができず，さらに，モーションの違いによる印象変化には対応していない．

モーション情報を用いた研究としては，沼口ら [5] によるダンスモーションを印象に基づき分類する研究があげられる．この研究では， Laban ら [6] が構築したラバン身体動作表現理論（Laban Movement Analysis，以下 LMA）に基づくモーション情報とユーザーテストから得られた印象評価とを比較し，それらの関連性を明らかにした．しかし，こ

の研究では楽曲の印象を考慮していない．

そこで，本研究では従来研究 [2,4] を基に，音色やリズムなどの音響特徴量と LMA に基づいたモーション情報を用いて重回帰分析を行うことで，VA 平面上での印象軌跡を求め，その印象に対応するキャラクタの表情を生成する．また，歌詞のない楽曲にも対応できるようにするため，本研究では音響特徴とモーション情報のみを用いることとする．

3. 提案手法

本研究では重回帰分析を用いて，音響特徴量とモーション特徴量から VA 平面上の軌跡として印象の軌跡を推定する．ここで，VA 平面とは，興奮 - 弛緩を表す Arousal（活発度）の軸および，快 - 不快を表す Valance（感情価）の軸からなる感情平面である．以下，具体的な手法について説明する．

3.1 重回帰分析による印象推定

音響特徴量とモーション特徴量を説明変数，それらに対応する 1 秒毎の VA 座標値を目的変数とする，以下の式(1)で表される式で重回帰分析を行う．

$$V = a_0 + \sum_{i=1}^{N} a_i x_i , \quad A = b_0 + \sum_{i=1}^{N} b_i x_i \qquad (1)$$

式(1)において，x_i は音響特徴量およびモーション特徴量，a_i, b_i は重回帰分析の偏回帰係数である．後述の主観評価実験を行うことで取得した VA 座標値を正解値データとして用いることで，最適な重回帰式の偏回帰係数を求める．ここで, 説明変数として使用する特徴量について説明する．前章（2.2 節）で述べたラバンの身体動作表現理論によるとモーションの印象を表す要素は「空間性」と「力動性」の 2 種類に大別することができる．そこで，我々はモーション特徴量として, 沼口ら [5] が用いた Laban の身体理論に関する特徴量を参考に, 表 1 の 13 次元の特徴量を選択した．また，音響特徴量に関しては既存研究[1]を基にリズム，音色に関する特徴量 80 次元を用いた（表 2）．特徴量を抽出する際は，MIR toolbox Ver.1.3 [7] を用い，得られた特徴量に対し以下の式(2)で正規化を行った．ここで，fは正規化前の特徴量，$\bar{f}$，f_σはそれぞれ，正規化前の特徴量平均，分散である．

3.2 正解値データ作成

我々は，主観評価実験を行うことで上記の重回帰分析における正解値データを取得する．その際に用いた評価用のダンス動画の作成方法について説明する．まず，正解値データとして，音響特徴量の変化に対応した VA 座標値のデータに加え，モーション特徴量の変化に対応した VA 座標値のデータが必要である．そこで我々は，様々な楽曲の特徴量の変化に関するデータを取得し，それらの各楽曲に対してモーションの力動性・空間性が大きいものと小さいもの（計 4 種類ずつ）を付与した動画を用意することで，モーションに同期した印象変化データを得ることを考えた．

また，正当な印象評価を行うため，評価実験の際に用いるダンス動画は，楽曲とモーションとの間で印象やテンポの違いによる違和感をなくす必要がある．そこで我々は，楽曲とそれに対応するモーションを，予め楽曲の印象推定により VA 平面の各象限に対応する 4 つの印象にクラスタリングし，同一クラスタに属する楽曲とモーションとを組み合わせることでダンス動画を作成した．以下，評価実験で使用する動画の作成手順について示す．

表 1．使用するモーション特徴量

モーション特徴量	概要
空間性	ルートから両手足間の距離，両手足で囲まれた領域の面積（平均・分散）
力動性	全身の角速度，身体移動速度（平均・分散）

表 2．使用する音響特徴量

楽曲特徴量	概要
リズム	オンセット振幅，テンポ
音色	MFCC（20 次元），スペクトル重心，スペクトルフラックス，スペクトルロールオフ，スペクトルフラットネス，スペクトルコントラスト（8 音域）

$$f' = \frac{f - \bar{f}}{f_\sigma} \qquad (2)$$

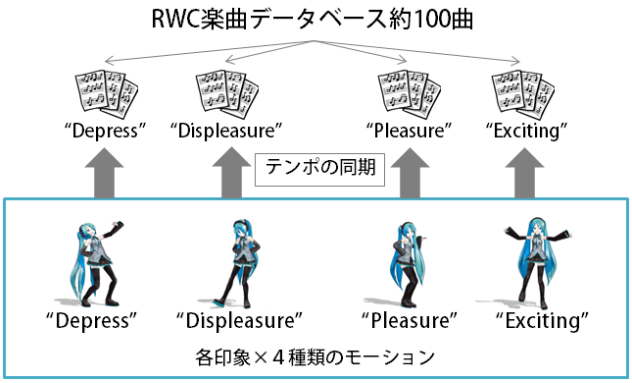

図 4．動画生成手順

i) 使用するモーションの選択

まず，いくつかのモーションセット（モーションと楽曲が予めセットになったもの）を用意する．これらのモーションセットは実際にアーティストが作成したもので，楽曲の印象にマッチしたモーションが付与されているため，楽曲とダンスの印象は同じであると仮定した．次に，モーションセットの各楽曲の印象を推定することで，VA 平面の第 1〜4 象限（4 種類の印象）に対応するモーションを 1 つずつ選択する．ここで，VA 平面の第 1 象限は Exciting，第 2 象限は Depress，第 3 象限は Displeasure，第 4 象限は Pleasure といった印象語に対応する．なお，楽曲の印象推定には重回帰分析を用い，正解値データとして，240 楽曲（各 15 秒間）に対し 1 秒毎に印象評価された VA 座標値が収録してある MoodSwing Turk Dataset [8,9] を使用した．

ii) 使用するモーションの切り出し

(i)で得られた 4 楽曲に対応するモーションの中で空間性・力動性の平均値が最大・最小となる領域をそれぞれ 20 秒間ずつ切り出し，4 印象の楽曲×4 種類のモーション＝合計 16 種類のモーションを取得する．ここで，空間性を測る特徴量としてルート位置と各関節間の距離の平均値，力動性を測る特徴量として各関節角速度の平均値を用いた．

iii) 楽曲とモーションの結合

動画で使用する楽曲は RWC 研究用音楽データベース[10] のポピュラー楽曲 100 曲であり，各楽曲のサビ部分の 20 秒間を切り出した．それら 100 曲分の楽曲素片に対して(i)と同様に重回帰分析を用いて印象推定を行い，4 つの印象のいずれかに分類する．同じ印象に分類された楽曲と(ii)で用意した 4 種類のモーションを組み合わせる際に，モーションを伸縮・拡張させることで，モーションと楽曲のテンポを同期させた．以上の手順で，全部で 400 個の評価用ダンス動画を作成した（図 4）．

3.3 主観評価実験；

本節では，作成した評価用ダンス動画を評価してもらい，正解値データを作成するための評価実験の方法について説明する．

まず，被験者に 1 人あたり 1 曲 20 秒間のダンス動画（25 曲分×各 4 種類の印象＝合計 100 個の動画）について印象を評価してもらう．今回は 20 代の男女合計 10 名の被験者に対して行った．また，評価実験を行うために図 6 のような GUI を用意した．この GUI では，画面左側で再生されるダンス動画を見ながら，画面右側の VA 平面上でマウスカーソルを移動させることで，1 フレーム毎の VA 座標の軌跡をトラッキングすることができる．ここで，データを統一するため，マウストラッキングは毎動画必ずマウスを VA 座標の原点の位置に置いてから開始することとした．また，動画の再生に際しては，仮に同じ楽曲 4 種類の動画を続けて再生した場合，モーションの違いに意識が集中してしまい，公平な印象評価を行うことができないと考えられる．そのため，評価用の動画はランダムに再生することとした．

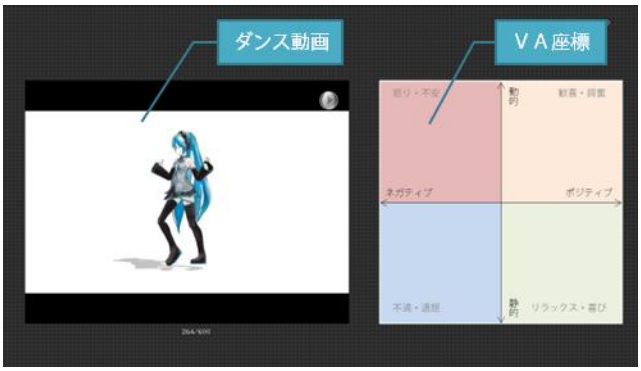

図 5. 評価実験で用いた GUI

3.4 表情アニメーション

本節では，印象推定結果をもとにダンスキャラクタに表情を付与する手法について説明する．

キャラクタの表情については，VA 平面上で Russell の感情語 [11] に対応する 8 つの表情を使用する．具体的には，図 6 に示すように VA 空間を 8 分割した際のそれぞれの領域に当てはまる感情語に対応する表情で，一番上から時計回りに「驚いた」，「いきいきした」，「嬉しい」，「のんびりした」，「眠い」，「悲しい」，「恐ろしい」，「怒った」の 8 つである．これらの感情語は「arousal」，「excitement」，「pleasure」，「relaxation」，「sleepiness」，「depression」，「displeasure」，「distress」を日本語に訳した表現 [12] である．キャラクタに付与する 8 つの表情はそれぞれ FACS [13] に基づいて作成した．FACS は「解剖学的に独立し，視覚的に識別可能な表情動作の最小単位」と定義される AU（Action Unit）を組み合わせて表情を定量的に記述したもので，これらの 44 種類の AU を組み合わせることで，人間

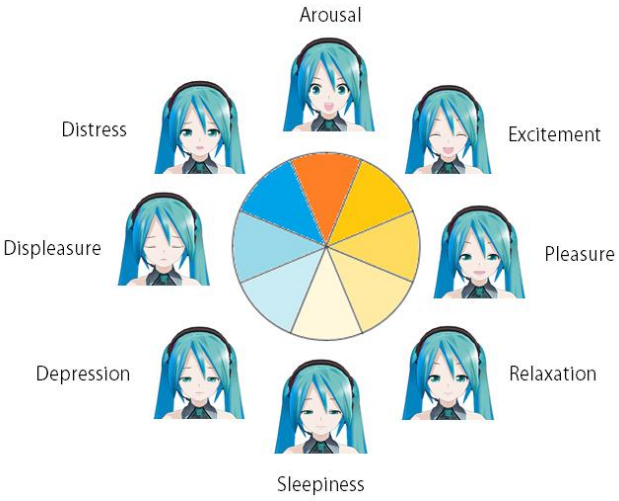

図 6. VA 平面と表情の対応[11]

のあらゆる表情を表すことができるとされている．

また，VA 平面の性質から原点に近い程それぞれの感情の度合いは小さくなるため，8 つの感情の度合いが最大のときに各感情の表情パラメタを 1, 感情の度合いが最小（VA 平面の原点）のときに各感情の表情パラメタを 0（無表情）とした．さらに，VA 座標値がある 2 つの感情の間に位置する場合は，その 2 つの感情からの距離を基に，その 2 つの感情に対応する表情を線形的にブレンドした表情を割り当てる．

4. 実験結果・考察

本章では，評価実験によって得た印象評価データをもとに精度評価した結果や考察について説明する．

4.1 特徴量の考察

重回帰式の説明変数として使用した特徴量に対して主成分分析をすることで，印象推定に有意な特徴量について考察する．

表 2 に第 1 主成分から第 3 主成分までで，寄与率の高いものから順に 93 次元の特徴量のうち上位 15%の特徴量をまとめたものを示す．これによると，第 1 主成分に寄与する特徴量として，間接間距離や間接角速度など，主にモーション特徴量が多く選択されている．第 2，第 3 主成分に関してはスペクトルコントラスト，MFCC が選択されている. また, 全体としては, 寄与率の高い特徴量として MFCC, 空間性（関節間距離），テンポなどが挙げられた．よって，作品の印象には楽曲中の音声やダンスキャラクタのポーズ, 楽曲のテンポなどが大きく寄与していることが分かった．

表 2．各主成分における特徴量寄与率

	第 1 主成分	第 2 主成分	第 3 主成分
寄与率(%)	8.685	8.244	6.360
寄与する特徴量	MFCC, ロールオフ, 関節間距離(最小・平均・最大), 面積, 移動速度(平均・最大・最小),	スペクトルコントラスト, MFCC	MFCC, スペクトルコントラスト

4.2 印象推定

主観評価実験から得た正解値データを用いて重回帰分析を行った結果を表 3 に示す．また，決定係数は使用する説明変数の個数に依存しない，自由度修正決定係数を用いた．決定係数が上がっていることから，ダンス動画の印象推定を行う際，音響特徴量だけでなくモーション特徴量も考慮することで, 印象推定精度が向上することが分かった．また，A 座標の決定係数は，V 座標のものに比べて大幅に低かったことから，求めた重回帰式は印象の激しさの度合いよりも，印象のポジティブ度合の推定に長けているといえる．この原因として考えられるのは， V 座標を決定する重回帰式と A 座標を決定する重回帰式において同一の特徴量を使用しているため，A 座標の印象推定に不利な特徴量が含まれているということである．

また，VA 座標の軌跡とそれに対応するモーション特徴量変化との類似度を計算したところ，最も類似度が高いモーション特徴量としては V 座標，A 座標共に関節間距離，次いで関節角速度が選ばれた．図 7 はある楽曲の印象推定結果の VA 座標値およびモーション特徴量の時間変化を表したもので，一番上からそれぞれ V 座標，A 座標，モーション特徴量(左から間接角速度，間接間距離)の時間変化を表している．図 7 からも分かるように，モーション特徴量の変動と VA 座標の軌跡の間に相関があると考えられる．しかし，それぞれのダンス動画によってその相関の傾向や度合いが異なっており, 普遍的な規則性は見られなかった．

表 3．楽曲特徴量からの印象推定と本手法の比較

印象推定法	決定係数（V 座標）	決定係数（A 座標）
本手法	0.8081	0.3582
音響特徴量のみ	0.7253	0.2788

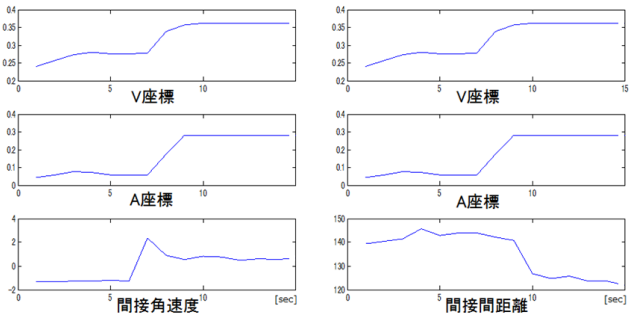

図 7．楽曲の VA 座標とモーション特徴量の例

4.3 表情アニメーション

次に，VA 平面の楽曲の印象軌跡からダンスキャラクタの表情アニメーションを生成した．図 8 はある楽曲の印象軌跡で，楽曲のみを用いて印象推定した場合と比較を示したものである．この図からは，モーション情報を用いることによる印象の大きな変動は見られないが，モーションの変化に追従して印象軌跡が少しずつ変化していることが分かる. 例えば, 楽曲は元気な印象でもモーションは動かず, ポーズをとっているだけの場合，図 9 のように，音響特徴量のみの場合に比べて A 軸（激しさ）の度合が低くなった．また，音響特徴量に大きな変化がないにも関わらず，モーションの力動性が大きくなると，A 軸の値が上がり，より元気な印象の表情が生成される傾向がみられた．

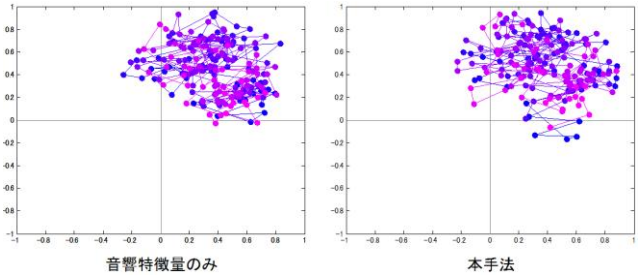

図 8. ある楽曲の印象軌跡比較（時間変化：青→ピンク）

図 9. 表情生成結果比較

5. おわりに

我々は，キャラクタの表情を自動合成するためのダンスモーションに同期した楽曲印象推定手法を提案した．音響特徴量とモーション情報によるダンス動画に特化した印象推定を行うことで，キャラクタの適切な表情を自動生成することが可能となった．また，モーション情報を考慮することで作品の印象推定精度が上がり，モーションの変化にマッチした表情変化が実現できた．

今後の課題としては，本手法で生成した表情アニメーションに対する主観評価実験や，VA 平面上の軌跡に対応する表情のブレンド方法の再検討があげられる．また，楽曲の歌詞や使用するキャラクタ，ステージなどの要素を考慮した印象推定にも取り組みたい．

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．

参考文献

[1] Thomas, L., Clarissa, M., João, P. Scarlett, B., Allyson, C., Renata, N., Vinícius, M., Adelailson, P., Dimas M., Thales, V., “Stereo Music Visualization through Manifold Harmonics,” Vis Comput, 2011.

[2] DiPaola, S. and Arya, A. “Emotional Remapping Of Music to Facial Animation.” In Proceedings of ACM SIGGRAPH symposium on Videogames, 143-149, 2006.

[3] Russall, J.A. “A Clrcumplex Model of Affect,” Journal of Personality and Social Psychology, 39(6), 1161-1178, 1980.

[4] 西川直毅，糸山克寿，藤原弘将，後藤真孝，尾形哲也，奥乃博，″歌詞と音響特徴量を用いた楽曲印象軌跡推定法の設計と評価．″ 第 91 回音楽情報科学研究会，91(7), 1-8. 2011.

[5] 沼口 直紀，中澤 篤志，竹村 治雄，“印象語による舞踊動作データの分類法．” 情報処理学会 コンピュータビジョンとイメージメディア研究会(CVIM)，167(35), 1-6，2009.

[6] Laban, R. and Ullmann, L. “Mastery of Movement,” Princeton Book Company Publishers, 1960.

[7] O. Lartillot, and P. Toiviainen, “MIR in matlab (II): A Toolbox for Musical Feature Extraction from Audio,”Proceedings of the 5th International Conference on Music Information Retrieval, pp. 127-130, 2007.

[8] Speck, J.A., Schmidt, E.M., Morton, B.G. and Kim, Y.E. ”A Comparative Study of Collaborative vs. Traditional Annotation Methods.” In Proceedings of the 12th International Society for Music Information Retrieval Conference(ISMIR), 549-554, 2011.

[9] Schmidt, E.M. and Kim, Y.E. “Modeling Musical Emotion Dynamics with Conditional Random Fields,” In Proceedings of the 12th International Society for Music Information Retrieval Conference(ISMIR), 777-782, 2011.

[10] Goto, M., Hashiguchi, H. , Nishimura, T. and Oka, R. “RWC Music Database: Popular, Classical, and Jazz Music Databases.” In Proceedings of the 3rd International Society for Music Information Retrieval Conference (ISMIR), 287-288, 2002.

[11] James, A.R., “A Circumplex Model of Affect, ” Journal of Personality and Social Psychology,Vol 39, No.6, 1161-1178, 1980.

[12] 菊谷麻美，小川時洋，鈴木直人，“感情語の 2 次元空間内の布置について．” 同志社心理，1998.

[13] Ekman, Paul, Wallace V. Friesen, and J. Hager. “The Facial Action Coding System (FACS): A Technique for the Measurement of Facial Action. Palo Alto.” CA: Consulting Psychologists Press, 1978.