Vol.2014-MUS-104 No.14
2014/8/26

MachineDancing: ダンス動作データの自動分析に基づく音楽に連動したダンス生成手法

深山 覚[1]　後藤 真孝[1]

概要：ダンスと音楽が連動したデータから機械学習を行い、新しく入力する楽曲に対して 3 次元コンピュータグラフィックスのキャラクタのダンスを自動生成できる手法 MachineDancing を提案する。従来、ダンスの断片を準備しそれを確率モデルなどを用いて音楽に合わせて接続することでダンス自動生成が実現されてきた。しかしダンスの断片を切り貼りするのみで、ダンス動作自体の学習・生成手法とはなっておらず、生成結果のバリエーションに限界があった。本研究ではダンス動作の確率モデルとしてガウシアンプロセス（GP）を用い、ダンスと音楽の対応関係のみでなく、ダンス動作自体をも学習することで、新たな動作を楽曲に連動して自動生成できる手法を提案する。

1. はじめに

3 次元（3D）コンピュータグラフィックス（CG）のキャラクタは広く普及したが、その動作の制作は人手によるものが多い。キャラクタの状況に応じて多様な動作をつくるには人手では限界があり、動的に自動生成する必要がある。そこで本研究では、音楽に連動したダンスを対象に、機械学習に基づいて新たな動作を自動生成できる手法を実現することを目的とする。

音楽に連動した 3D の CG キャラクタの動作を編集できるソフトウェア MikuMikuDance (MMD) [*1] により、ダンス動画は MMD の登場以前より手軽に作成できるようになった。MMD で用いることができるキャラクタのモデル (姿形や関節の構造など) はインターネット上に多く公開されており、自分自身でモデルをデザインできない人でもモデルを公開する人のおかげでアニメーションを作成できる。またキーフレームと呼ばれる特定の時刻でのキャラクタの姿勢 (ポーズ) を編集し、それらキーフレームのポーズ間の補間方法を設定するだけで、一連のキャラクタの動作を作成することができる。

しかしダンスを制作することは依然大変な作業である。第一に音楽に対応したダンス動作を制作するのが難しいためである。動きが単調にならないよう十分なバリエーションの動作を考える必要がある。また 3D キャラクタに対して不自然な動作が容易に設定できてしまうため、自然になるよう配慮して制作する必要がある。第二に、ダンスを入力することに時間がかかるためである。自然な動きにするためには、一曲を通じて多数のキーフレームを設定する必要があり、しかもそれらの補間曲線を試行錯誤しながら決めなければならない。少ない操作で自然な動作とするため、センサデバイス Kinect などによる人間の姿勢情報の取得 (モーションキャプチャ) を利用することも可能であるが、ダンスが踊れなければ用いることができない。

本研究ではこれらの困難を克服できる、音楽に連動したダンスを自動生成できる手法 MachineDancing を提案する。楽曲とダンスが対応付いた学習データを用いて、ダンス動作として自然な動作の機械学習を行うことで、任意の

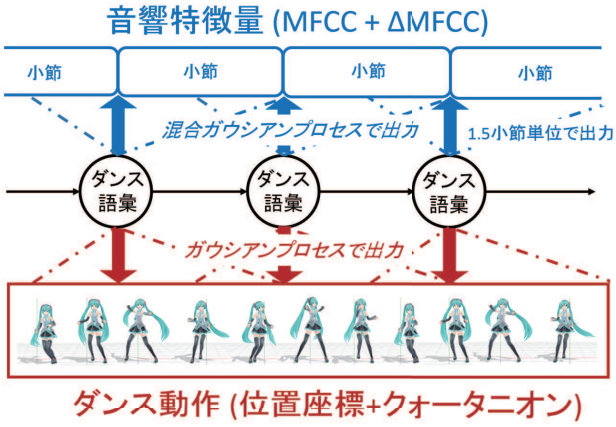

図 1 MachineDancing の確率モデル: ダンス語彙からガウシアンプロセスによってダンス動作と音響特徴量が出力される。ダンスと楽曲の対応付いたデータベースから機械学習を行い、任意の楽曲に対してダンスを自動生成することができる。

[1] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

[*1] http://www.geocities.jp/higuchuu4

楽曲に対してダンスを自動生成できる。なおダンス動作を切り貼りする既存の手法とは異なり、学習データのダンスを単に出力するのではなく、楽曲に応じた新たな動作を生成することができる。また学習データを大きく変えれば、それに伴って大きく異なったダンスを生成できる。

音楽に連動したダンスを自動生成する従来の研究では、音楽のテンポ [1][13]、ビート検出のための特徴量 [5][14]、音高と和音 [11]、メロディの概形 [12] などがダンスと対応づけて分析され用いられている。また 2 つ以上の音響特徴量を組み合わせて用いる手法 [3] や、1 曲中の音響特徴量の類似度行列を用いる手法 [8] がある。隠れマルコフモデル（HMM）を用いてダンスと音楽の対応関係をモデル化しダンス生成を行う手法も提案されている [11]。

一方でダンス動作の時系列を分析し、時間的に離れたキーフレーム間の自然な補間や、人間らしい動きを実現する研究も行われている。HMM [16]、動的ベイジアンネットワーク (DBN) [7]、階層的ディリクレ過程隠れマルコフモデル (HDP-HMM) [17]、Kernel Canonical Correlation Analysis (KCCA) [4]、ガウシアンプロセス (GP) を使う方法 [9]、潜在変数の軌跡のトポロジーを考慮した Topological Gesture Analysis (TGA) [10][15] が提案されている。またダンス動作ではなく人間の動作一般の研究として、Gaussian Process Dynamical Models (GPDM) [18]、Multi-layer Joint Gait-Pose Manifolds (multi-layer JGPM) [2] などを用いて、人間の動作を低次元の空間へ非線形写像によって縮退させて分析・補間する手法が提案されている。

このように従来の方法では楽曲に連動したダンスが生成されるものの、ダンス動作の時系列を分析し学習データにない新たなダンス動作を自動生成するには至っていない。またダンスや人間の動作の自然な補間を行うために GP を使う手法が提案されているが、それを用いて楽曲に対してダンスを自動生成できなかった。そこで MachineDancing では

1. HMM によるダンスと音楽の対応関係のモデル化
2. GP によるダンス動作と音楽特徴量のモデル化

により、楽曲とダンスが対応付いた学習データを用いて機械学習を行い楽曲に対して新しいダンス動作を生成する。

手法の概略を図 1 に示す。MachineDancing で用いられる確率モデルは、音楽の音響特徴量を観測系列とし、類似したダンス同士をまとめた「ダンス語彙」が隠れ状態となる HMM である。音楽の特徴量の出力確率はガウシアンプロセスによって決定され、またダンス動作もガウシアンプロセスによって各隠れ状態から確率的に生成される。ダンスの自動生成は、楽曲の音響特徴量を観測とした場合の尤度最大のダンス動作を探索する問題として定式化できる。

2. 音楽に連動するダンスの確率的生成モデル

本節では音楽に連動するダンスの確率的生成モデルを設計し、それを使って楽曲に連動したダンス動作を自動生成する方法を述べる。はじめに音楽とダンスをそれぞれ扱うための特徴量について述べる。次にダンス動作の確率的生成モデルを設計し、データに基づくそのモデルの学習アルゴリズムを述べる。さらに音楽をダンス動作の確率的生成モデルと対応付けるための方法を議論し、最後にそのモデルを用いて新たな楽曲に対してダンスを自動生成する方法を述べる。なお本研究では楽曲のビートと小節線が、階層的なビート構造の推定 [19] 等によって適切に与えられるものとする。

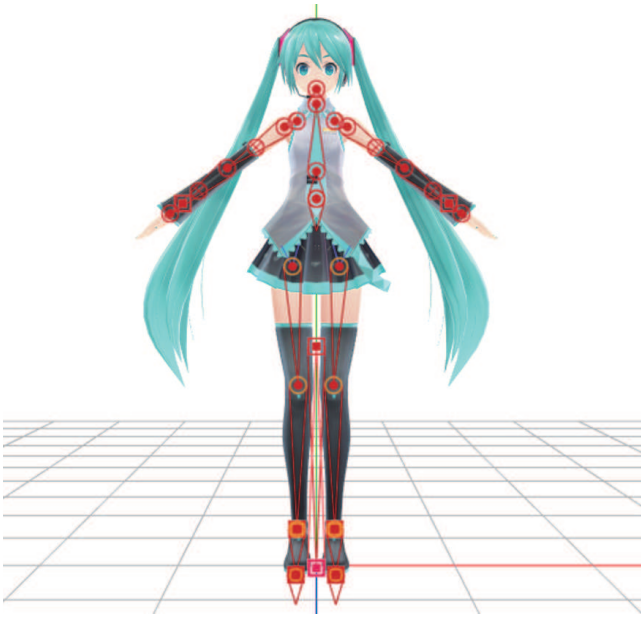

図 2 本手法でダンスを分析・生成する際に用いる部位（ボーン）。体の中心を表わすボーンはワールド座標によって表わし、他のボーンは接続しているボーンとの接点を中心とする回転を表現するクォータニオンによって表現される。図中のキャラクタの 3D モデルは koron 氏によって制作された。

2.1 ダンス生成のための音響特徴量とダンスのデータ構造

2.1.1 ダンス生成のための音響特徴量

音楽音響信号を分析するために広く用いられているメル周波数ケプストラム係数 （MFCC）（16 次元）と、直前の分析区間の MFCC との差分である ΔMFCC（16 次元）を連ねた 32 次元のベクトルを音響特徴量として用いる。これにより曲中の盛り上がり、類似部分（繰り返されるサビなど）、音色の判別をある程度行うことができる。しかしこれらは表層的な特徴量であるため、ダンスの動作とより密接に関係する特徴量は、今後検討の余地がある。

2.1.2 ダンス動作のデータ構造

3 次元キャラクタのモデルにはボーンと呼ばれる人間の骨格に相当する構造があり、このボーンを平行移動もしくは回転させることでキャラクタを動かすことができる。体の中心を表わすボーンの位置は動きに依存せず固定されたデカルト座標に基づく位置座標 (x_1, x_2, x_3) と表現できる。

しかし腕や足といったボーンは、他のボーンと接続されているため、位置座標で表すと腕と胴体が離れるなど姿勢として不可能なものが表現されてしまい、ふさわしくない。これらのボーンについては、体の中心に近い方へ接続されているボーン（親ボーン）との接点を中心として、基本姿勢（図 2 の姿勢）からどれだけ 3 次元的に回転されたかで表現できる。

3 次元の回転は回転行列による表現、オイラー角による表現、クォータニオンによる表現があるが、本研究ではクォータニオンを用いる。回転行列は 3×3 の行列であり、単位ベクトルの向きの変化を線形変換で表したときの変換行列である。オイラー角は、3 つの軸を事前に決めた順番で中心軸として回転させる角度で 3 次元回転を表現する方法で $(\theta_x, \theta_y, \theta_z)$ と表せる。クォータニオン Q は回転前後のベクトルの外積 (q_1, q_2, q_3) と内積 q_0 からなる四元数であり、$Q = q_0 + q_1\mathbf{i} + q_2\mathbf{j} + q_3\mathbf{k}$ と表わせる。ここで $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1$、$\mathbf{ij} = -\mathbf{ji} = \mathbf{k}$、$\mathbf{jk} = -\mathbf{kj} = \mathbf{i}$、$\mathbf{ki} = -\mathbf{ik} = \mathbf{j}$ である。クォータニオンでは 4 つのパラメータによって姿勢の回転を表わすことができる。

クォータニオンは回転行列よりパラメータ数が少なく、またオイラー角の表現の問題である、回転する軸の順番と角度によっては目的の姿勢に導く表現が存在しない場合や、姿勢間を補間する際に大回りとなる軌跡を描いてしまうジンバルロック現象を避けることができる。さらに球面線形補間が簡単に行えるなどの便利な性質がある。

本研究では体の動きを表現するために必要な腕、足などの 20 本のボーンを用いた。用いたボーンを図 2 に示す。全 20 本のボーンのうち、体の中心を表わす 3 次元座標の 3 つのパラメータと、残り 19 本のボーンの回転を表わすクォータニオンの 4 つのパラメータを 1 つのベクトルへと束ね、79 次元のベクトルとして各時刻でのキャラクタの姿勢を表現することができる。

2.2 ダンス動作の確率的生成モデル

2.2.1 ダンス語彙

楽曲に連動したダンス動作は部分に切り分けて捉えることができる。実際、サビや A メロの繰り返しや音響特徴が似ている部分では類似した動作が見られることが多い。またより小さな時間単位においても、2 つの姿勢が交互に現われるなど繰り返し構造がダンス動作の中に見られる。繰り返されている部分は、その場所の音楽や、それが何回めの繰り返しかに応じて、変化して現われることが多いが、それらが類似しているため、同じ動作から派生して生まれた動作として捉えることができる。

このような類似したダンス動作を生成する基本形となるダンス動作のことを、本研究では「ダンス語彙」と呼ぶ。楽曲中の A メロの繰り返しにおいて、類似したダンス動作がある場合、それらは類似した共通したダンス語彙から生成されていると捉えることができる。

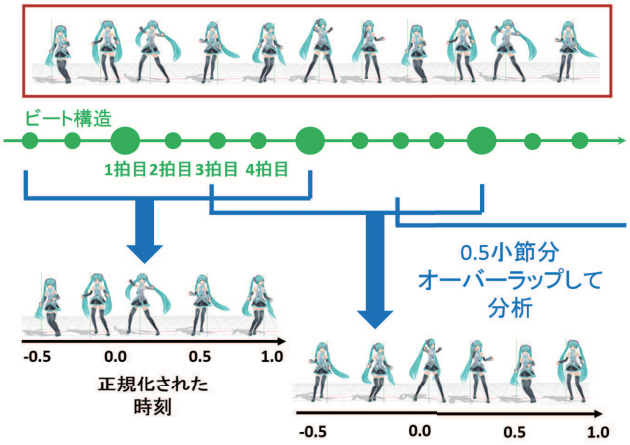

図 3 ダンス動作の分析区間への切り出し方: 0.5 拍のオーバーラップをさせながら 1 小節間隔で動作を分析する。テンポの異なる楽曲のダンスから切り出した動作同士で比較をするため、1 小節を 1.0 とするよう正規化する。

本研究では楽曲に連動したダンス動作を分析区間へと分割して捉え、分析区間ごとにダンス動作がダンス語彙から確率的に生成されている、という発想で確率的生成モデルを設計する。

ダンス語彙はどのような時間的長さで定義されるべきだろうか。音楽に連動したダンスは音楽の拍の位置に大きく関連しており、時間の単位として、1 拍の長さを規準にするのがふさわしい。一般に分析区間が長いほど長い時間的構造を持ったダンス語彙を分析できるが、語彙数が膨大になってしまう欠点がある。本研究では 4 分の 4 拍子を仮定し、1 小節 4 拍の中での強拍と弱拍でどのようにダンス動作が違うのかについて観察できる区分の長さとするため、1 小節のダンスを基本単位と考える。その上で新たな提案として、小節の先頭の動作へ前の小節から続く動作も合わせて分析するために、前の小節の 3 拍目から始まり、その小節の 4 拍め最後で終る 1.5 小節の分析区間を用いてダンスを分析する。ダンス動作の分析区間の様子を図 3 に示す。このようにダンス動作が 0.5 小節分オーバーラップしながら切り出され分析する方法を提案する。なお、分析するダンスに対応した楽曲によってテンポが異なるため、切り出されるダンスの時間長が異なり、互いに比較するのが不便である。そこで 1 小節が 1.0 となるように時間の正規化を行う。

2.2.2 ダンス語彙に基づくダンス動作の確率的生成モデル

先頭から n 番目の分析区間のダンス $\mathbf{D}_n$ が、ダンス語彙 v_n として分析することができるとする。ここで $\mathbf{D}_n$ は分析区間中の姿勢を表わすベクトル $\mathbf{d}$ を束ねた行列 $[\mathbf{d}_1, \cdots, \mathbf{d}_s, \cdots, \mathbf{d}_S]^T$ である。$\mathbf{d}_s$ は現在の実装では 79 次元ベクトルである。S は分析区間中で観測される姿勢の個数である。ダンス動作のデータには一定のフレームレート

で時刻ごとの姿勢が記述されるか、時間的にとびとびのキーフレームの時刻ごとの姿勢が記述されている。したがって、分析区間中に含まれる姿勢データの個数 S は、楽曲のテンポやキーフレームの詳細な設定され度合いに応じて異なっている可能性がある。

分析区間内のダンス動作 $\mathbf{D}_n$ がダンス語彙 v_n を元に確率的な揺らぎをもって生成されるとして、確率 $P(\mathbf{D}_n|v_n)$ を導入する。またダンス語彙は隣合う分析区間にランダムに連なるわけではなく、ある語彙の後にはどのような語彙が続きそうかという偏りがあると考えられる。そこで語彙から語彙へと遷移する確率 $P(v_n|v_{n-1})$ を導入する。

このとき N 個の分析区間のダンス動作 $\{\mathbf{D}_n\}_{n=1}^N$ がダンス語彙 $\{v_n\}_{n=1}^N$ から生成される確率は以下のように計算できる。

$$P\left(\{\mathbf{D}_n\}_{n=1}^N, \{v_n\}_{n=1}^N\right) = \prod_{n=1}^N P(\mathbf{D}_n|v_n) P(v_n|v_{n-1}) \quad (1)$$

ただし、表記の便宜上 $P(v_1|v_0) = P(v_1)$ とした。以後の式でもインデックスが 0 となる場合は同様に扱う。これは HMM であり、観測がダンス動作、隠れ状態 (潜在変数) がダンス語彙である。$P(\mathbf{D}_n|v_n)$ を出力確率、$P(v_n|v_{n-1})$ は遷移確率、$P(v_1)$ を初期確率と呼ぶ。ここで $P(\mathbf{D}_n|v_n)$ はダンス語彙が与えられたときにどのようなダンス動作が出力されるかについての確率モデルであり、これを「ダンス語彙からの確率的生成モデル」と呼ぶ。次項で設計されるこの確率的生成モデルを上記のモデルと統合することで、ダンス動作の確率的生成モデルをつくることができる。

2.2.3 ダンス語彙からの確率的生成モデル

ダンス動作には周期性や拍ごとの動作に対応関係があることから、これらを確率的にモデル化したい。同じダンス語彙として捉えることができる 2 つの似たダンス動作 $\mathbf{D}_1$ と $\mathbf{D}_2$ があるとき、これら 2 つの行列に含まれる姿勢を表わすベクトルの個数は同じとは限らない。そこで、$\mathbf{D}$ に含まれる姿勢 $[\mathbf{d}_1, \cdots, \mathbf{d}_S]$ を、1 小節の長さが 1.0 となるよう正規化された時刻を変数として、$\mathbf{d}_s = f(t_s)$ と表せる連続関数 f を推定する必要がある。ここで t_s は姿勢 $\mathbf{d}_s$ が分析区間内で観測される時刻 $(-0.5 < t_s < 1.0)$ である。f を推定するにあたって、次の 2 点を仮定する。

第一の仮定として、各時刻での姿勢 $\mathbf{d}_s$ は、$f(t_s)$ の値からガウスノイズが加わって観測されるとする。すなわち

$$|\mathbf{d}_s - f(t_s)|^2 \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right) \quad (2)$$

と書ける。ここで $\mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$ は平均が零ベクトルで、分散共分散行列が $\sigma^2\mathbf{I}$ である多次元ガウス分布である。

第二の仮定として、関数 $f(t)$ が、適切に選ばれた（非線形）基底関数 $\phi_j(t)$ の重み付け和で

$$f(t_s) = \sum_{j=1}^J a_j \phi_j(t_s) \quad (3)$$

と表現されたとき、重み係数 $\{a_j\}_{j=1}^J$ が平均 $\mathbf{0}$、分散共分散行列 Σ_p のガウス分布に従い、

$$\mathbf{a} = (a_1, \cdots, a_J) \sim \mathcal{N}(\mathbf{0}, \Sigma_p) \quad (4)$$

となるとする。このとき Representer 定理により、基底関数 ϕ_j が正定値カーネルであるときには、第一の仮定のもとで推定される f が式（3）の形を持ち、J は $\mathbf{D}$ に含まれる姿勢の数（行列の行数）となることが知られている。本研究では正定値カーネルとして RBF カーネル

$$\phi_{t_j}(t) = k(t, t_j) = \exp\left(-\frac{\lambda}{2}|t - t_j|^2\right) \quad (5)$$

を用いた。

これら 2 つの仮定のもと $P(\mathbf{D}, \mathbf{a}) = P(\mathbf{D}|\mathbf{a}) P(\mathbf{a})$ を計算し、この同時分布中の $\mathbf{a}$ を積分消去すると、分析区間のダンス動作の確率的生成モデルは

$$P(\mathbf{D}|\mathbf{t}; \sigma, \lambda) = \frac{1}{\sqrt{(2\pi)^{|\mathbf{t}|\alpha} |\mathbf{K}|^\alpha}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{D}^T\mathbf{K}^{-1}\mathbf{D}\right)\right) \quad (6)$$

と求まる。これはガウシアンプロセスによるモデル化である。ただし $\mathbf{t}$ は分析区間中で姿勢が観測される時刻を束ねたベクトル $\mathbf{t} = (t_1, \cdots, t_s, \cdots, t_S)^T$、$\alpha$ はベクトル $\mathbf{d}$ の次元数、$\mathbf{K}$ は RBF カーネルに基づくカーネル行列で $(\mathbf{K})_{ij} = k(t_i, t_j) + \sigma^{-2}\delta_{t_i t_j}$ であり、δ はクロネッカーのデルタである。

2.3 ダンス動作の確率的生成モデルの学習アルゴリズム

前節で設計されたダンス動作の確率的生成モデルを、データに基づいて学習する方法を議論する。

2.3.1 ダンス語彙の学習

ダンス語彙 $\{v_k\}_{k=1}^K$ を学習する方法を議論する。ダンスの語彙には自然言語のようなあらかじめ決まった語彙のセットはないので、ダンス動作の類似度に基づいてクラスタリングを行い学習する。

最も適切にダンス語彙とダンス動作の対応関係が取れている状態は、ダンス語彙が割り当てられたときの学習データの対数尤度 $L = \sum_{n=1}^N \log P(\mathbf{D}_n|\mathbf{t}_n, v^*; \sigma_{v_n}, \lambda_{v_n})$ が最大となるときといえる。この状態は、適当なダンス語彙とダンス動作の対応関係の初期状態から、EM アルゴリズムに基づくクラスタリングで求めることができる。

- **E-step**: 語彙とダンス動作の対応関係をもとに、各語彙 v ごとに σ_v と λ_v を求める
- **M-step**: データ全体の尤度が最大となるよう、語彙とダンス動作の対応関係を決める

の 2 つのステップの反復によって行える。以下、E-step と M-step のそれぞれの詳細を述べる。

2.3.1.1 E-step

M 個の分析区間のダンス動作 $\{\mathbf{D}^1, \cdots, \mathbf{D}^M\}$ がダンス

語彙 v^* であるとわかった場合、ダンス語彙 v^* が与えられたときのダンス動作の確率的生成モデルを式（6）に基づいて作ることができる。$\mathbf{D}$ に $\mathbf{D}^* = \left[\mathbf{D}^1, \cdots, \mathbf{D}^M\right]^T$ を代入し、$\mathbf{t}$ へ各分析区間において姿勢が観測される時刻の行列 $\mathbf{t}^* = \left[\mathbf{t}^1, \cdots, \mathbf{t}^M\right]^T$ を代入したときの、σ と λ を変数と見たときの値、すなわち尤度を最大化するように σ と λ を求めれば良い。具体的には対数尤度 $LL(\sigma, \lambda, v^*) = \log P(\mathbf{D}^*|\mathbf{t}^*, v^*; \sigma, \lambda)$ を最大化する σ_{v^*} と λ_{v^*} を、Scaled Conjugate Gradient [7] などの最適化法によって求める。この尤度は語彙 v^* の下で集められたダンス動作から学習して計算されるものであるので、条件に v^* が入っていることに注意されたい。

2.3.1.2 M-step

E-step で得られる語彙 $\{v_k\}_{k=1}^K$ に対応する確率的生成モデルがあれば、あるダンス動作 $\hat{\mathbf{D}}$ がどの語彙から生成されるものかを、対数尤度最大となる語彙の確率的生成モデルを探索することで調べることができる。すなわち、最も尤もらしい $\hat{\mathbf{D}}$ を生成する語彙 $\hat{v}$ は $\hat{v} = \mathrm{argmax}_v \log P\left(\hat{\mathbf{D}}|\hat{\mathbf{t}}, v; \sigma_v, \lambda_v\right)$ によって求めることができる。

2.3.2 出力確率・遷移確率・初期確率の学習

出力確率、すなわちダンス動作 $\mathbf{D}$ がダンス語彙 v から生成される確率は、最適化されたモデルパラメータ σ_v と λ_v を用いて、学習データが与えられたもとでの条件付き確率を以下のように計算すればよい。

$$P(\mathbf{D}|v, \mathbf{D}^*, \mathbf{t}^*, \mathbf{t}) = \frac{1}{\sqrt{(2\pi)^{|\mathbf{t}|\alpha}|\mathbf{K}|^\alpha}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{Z}^T\hat{\mathbf{K}}^{-1}\mathbf{Z}\right)\right) \quad (7)$$

ただし

$$\mathbf{Z} = \mathbf{D} - \mathbf{A}\mathbf{K}^{-1}\mathbf{D}^*, \quad \hat{\mathbf{K}} = \mathbf{B} - \mathbf{A}^T\mathbf{K}^{-1}\mathbf{A} \quad (8)$$

$$(\mathbf{A})_{ij} = \exp\left(-\frac{{\lambda_v}^2}{2}|t_i^* - t_j|^2\right) + {\sigma_v}^2\delta_{t_i^* t_j} \quad (9)$$

$$(\mathbf{B})_{ij} = \exp\left(-\frac{{\lambda_v}^2}{2}|t_i - t_j|^2\right) + {\sigma_v}^2\delta_{t_i t_j} \quad (10)$$

である。ここで式（7）の値は語彙 v の下で集められたダンス動作から学習して計算されるものであるので、条件に v が入っている。

初期確率と遷移確率は、学習データのすべてのダンス動作について語彙を推定した上で、語彙の出現回数、隣り合う語彙の遷移回数をもとに計算すればよい。

2.4 音楽に連動するダンス動作のモデル化

2.4.1 音楽とダンスの対応関係の確率モデル

ダンス動作の語彙 v のもとでの音楽音響信号を分析するため、音響特徴量 MFCC+ΔMFCC の列についても、ダンス動作の分析区間と同じように、一つ前の小節の 3 拍目から今の小節の 4 拍めまでを切り出して分析する。切り出された音響特徴量の列を $\{\mathbf{M}_n\}_{n=1}^N$ とする。

ダンス動作の議論と同様に、音響特徴量にも時間的な構造があり、小節中の特徴量ベクトル同士の相関がダンス動作に関連していると考えられる。そこで $P(\mathbf{M}|v)$ をガウシアンプロセスによってダンス動作の場合と同様にモデル化したい。しかし、同じダンスに対して対応する音響特徴量の列は多様であり、単一の確率分布で表現することはできないと考えられる。そこで複数の確率分布の重み付き和によって確率を計算することを考える。

あらかじめ $\{\mathbf{M}_n\}_{n=1}^N$ を、ダンス動作をクラスタリングした場合と同様の操作で、K 個のクラスタ $\{z_k\}_{k=1}^K$ に分割しておく。このとき $P(\mathbf{M}_n|v)$ は z を変数に加えて、

$$\begin{aligned} P(\mathbf{M}_n|v_n) &= \sum_z P(\mathbf{M}_n, z|v_n) \\ &= \sum_z P(\mathbf{M}_n|z)\, P(z|v_n) \end{aligned} \quad (11)$$

と書ける。$P(\mathbf{M}_n|z)$ は音響特徴のクラスタ z が与えられたときの音響特徴量の列の確率であり、これはクラスタリングによってすでに求まっているガウシアンプロセスに基づいて決定される。$P(z|v_n)$ は、ダンスの語彙が与えられたときの音響特徴量のクラスタが出現する確率であり、学習データにおいて、ダンス語彙ごとに、そのダンス語彙と同時にどのような音響特徴量のクラスタが観測されたかをカウントすることで求めることができる。

式（11）において、$P(\mathbf{M}_n|z)$ はガウシアンプロセスであり、ガウシアンプロセスに重み $P(z|v_n)$ $(\sum_z P(z|v_n) = 1)$ がかけられているため、これは混合ガウシアンプロセスとなっている。

2.4.2 ダンス動作の確率的生成モデルとの統合

あるダンス語彙 v のもとで、どのような MFCC+ΔMFCC が生成されるかの確率に基づいて、ダンス動作の確率的生成モデルと統合し、音楽に連動するダンス動作の確率的生成モデルをつくることができる。HMM である式 (1) にダンス語彙 $\{v_n\}_{n=1}^N$ を導入して変形すると、

$$\begin{aligned} &P\left(\{\mathbf{M}_n\}_{n=1}^N, \{\mathbf{D}_n\}_{n=1}^N\right) \\ &= \sum_{\{v_n\}_{n=1}^N} P\left(\{\mathbf{M}_n\}_{n=1}^N, \{\mathbf{D}_n\}_{n=1}^N, \{v_n\}_{n=1}^N\right) \\ &= \sum_{\{v_n\}_{n=1}^N} \prod_{n=1}^N P(\mathbf{M}_n|v_n)\, P(\mathbf{D}_n|v_n)\, P(v_n|v_{n-1}) \end{aligned} \quad (12)$$

を得て、ダンス動作のガウシアンプロセスに基づくモデルと音響特徴量とダンス語彙の関係のモデルを用いて確率が計算できることがわかる。

2.5 音楽に連動したダンスの自動生成

このように得られたモデル $P\left(\{\mathbf{M}_n\}_{n=1}^N, \{\mathbf{D}_n\}_{n=1}^N\right)$ をもとに、ビートと小節線が推定済みの新しい楽曲に対してダンス動作系列を生成する。ダンスの自動生成は与えられ

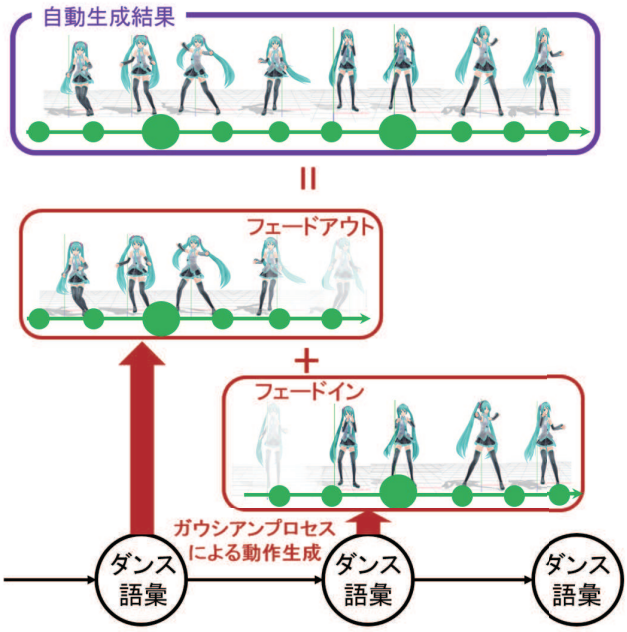

図 4 ダンス動作の自動生成: 各ダンス語彙からガウシアンプロセスによって生成される 1.5 小節単位のダンス動作は、3 拍目 4 拍目においてフェードイン・フェードアウトするように加算され、楽曲を通じたダンス動作が生成される。上の図中ではキャラクタの姿がフェードイン・フェードアウトしているが、実際にはキャラクタのボーンの位置・回転が線形補間される。

た音楽特徴量のもとで尤度最大のダンスを探索する問題として帰着できる。しかし各時刻でのダンス動作を最適化することは困難であるので、ダンス語彙の列 $\{v_n\}_{n=1}^N$ を最適化する問題として解き、得られたダンス語彙の列からダンス動作を生成する 2 ステップによって自動生成を行う。

まずはじめに、分析区間ごとに与えられた音響特徴量 $\{\mathbf{M}_n\}_{n=1}^N$ のもとで、尤度最大のダンス語彙の系列 $\{v_n^*\}_{n=1}^N$ を求める。

$$\{v_n^*\}_{n=1}^N = \underset{\{v_n\}_{n=1}^N}{\operatorname{argmax}} \prod_{n=1}^{N} P\left(\mathbf{M}_n|v_n\right) P\left(v_n|v_{n-1}\right) \quad (13)$$

これは音響特徴量が観測、ダンス語彙が隠れ状態である HMM であり、Viterbi アルゴリズムを用いて解くことができる。

次に得られたダンス語彙の列 $\{v_n^*\}_{n=1}^N$ を用いて、分析区間ごとにダンス動作 $\mathbf{D}_n^{\text{new}}$ を生成する。ダンス語彙 v_n としてクラスタリングされたダンス動作を集めた行列 $\mathbf{D}^* = \left[\mathbf{D}^1, \cdots, \mathbf{D}^M\right]^T$ を用い、ガウシアンプロセスの各時刻での平均を求めることで生成できる。また、ダンス動作が確率モデルによってモデル化されていることで、学習で得られた分散に基づいて平均値に擾乱を与えて、生成する度に異なったダンス動作を生成することも可能である。最後に、各分析区間でダンス語彙から生成されたダンス動作は、各小節の 3 拍め 4 拍目で前後の動きがフェードアウト・インするように、線形補間を行い、楽曲全体に連動するダンスを生成する。補間の様子を図 4 に示す。

3. ダンスの自動生成実験

本手法によって実際にダンス動作が楽曲から生成できるかを確かめた。また自動生成結果と学習データに含まれるダンス動作を比較し、学習データのダンスに基づいて新しい動作を自動生成できていることを確認できた。生成結果はホームページ `https://staff.aist.go.jp/s.fukayama/MachineDancing/index-j.html` から確認できる。

60 曲のダンスと楽曲が対応付いたデータを用いモデルの学習を行った。ダンスのデータは MMD のモーションデータフォーマット vmd によって記述されている。楽曲のビートと小節線は、階層的なビート構造の推定 [19] によって求めた後に、その推定誤りを手作業で修正した。0.5 小節ずつオーバーラップをさせながら、1.5 小節の分析区間長でダンス動作を切り出し、モデルの学習を行った。生成されたダンスは、3D の様々な CG キャラクタ（例えば、図 2 の koron 氏によって作成された初音ミクの 3D キャラクタ）を用いレンダリングした。

4. おわりに

本稿では楽曲とダンスが対応付いた学習データを用いて機械学習を行い、任意の新しい楽曲に連動したダンス動作を自動生成する手法 MachineDancing を提案した。学習データ中のダンス動作の断片を切り貼りするのではなく、音楽とダンスの対応関係に加えて、ダンス動作自体をも学習できる確率モデルを設計し、その学習アルゴリズムを導出した。またダンス語彙という概念を導入し、時間を変数とする連続関数によってダンス動作をモデル化したことで、異なるテンポの楽曲などを含む多様な学習データを用いて確率モデルを学習できるようになった。

現在の MachineDancing では、楽曲を与えると完全に自動でダンスを生成する手法だが、今後本手法を応用してユーザがインタラクティブにダンスを制作できるインターフェースを構築する予定である。

謝辞 本論文の図中に登場する 3D キャラクタは、ピアプロ・キャラクター・ライセンスに基づいてクリプトン・フューチャー・メディア株式会社のキャラクター「初音ミク」を使用しました。またキャラクタの 3D モデルは koron 氏によって制作されたものです。本研究の一部は JST CREST「OngaCREST」プロジェクトの支援を受けました。

参考文献

[1] K. M. Chen, S. T. Shen and S. D. Prior: “Using music and motion analysis to construct 3D animations and visualisations,” *Digital Creativity* Vol. 19, No. 2 (2008).
[2] M. Ding and G. Fan: “Multi-layer joint gait-pose man-

ifold for human motion modeling," In *Proc. FG 2013* pp. 1–8, (2013).

[3] R. Fan, S. Xu and W. Geng: "Example-based automatic music-driven conventional dance motion synthesis," *IEEE Transactions on Visualization and Computer Graphics* Vol. 18, No. 3 (2012).

[4] T. Hirose and T. Taniguchi: "Abstraction multimodal low-dimensional representation from high-dimensional posture information and visual images," *Journal of Robotics and Mechatronics* Vol. 25, No. 1 (2013).

[5] J. W. Kim, H. Fouad, J. L. Sibert and J. K. Hahn: "Perceptually motivated automatic dance motion generation for music," *Computer Animation and Virtual Worlds 2009* Vol. 20 (2009), pp. 375-384.

[6] M. Lee, L. Lee and J. Park: "Music similarity-based approach to generating dance motion sequence," *Multimedia Tools and Applications* Vol. 62, No. 3 (2013), pp. 895-912.

[7] M. F. Møller: "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks* 6, 4 (1993), pp. 525-533.

[8] K. Moon and V. Pavlović: "Graphical models for human motion modelling. Human Motion," *Computational Imaging and Vision* Vol. 36 (2008), pp. 159-183.

[9] T. Mukai and S. Kuriyama: "Geostatistical Motion Interpolation," In *Proc. ACM SIGGRAPH*, Vol. 24, No. 3 (2005), pp. 1062-1070.

[10] L. Naveda and M. Leman: "The spatiotemporal representation of dance and music gestures using topological gesture analysis (TGA)," *Music Perception* Vol. 28, No. 1 (2010), pp. 93-111.

[11] F. Ofli, E. Erzin, Y. Yemez and A. M. Tekalp: "Learn2dance: Learning statistical music-to-dance mappings for choreography synthesis," *IEEE Transactions on Multimedia* Vol. 14, No. 3 (2012).

[12] S. Oore and Y. Akiyama: "Learning to synthesize arm motion to music by example," In *Proc. WSCG 2006* (2006), pp. 201-208.

[13] C. Panagiotakis, A. Holzapfel, D. Michel and A. Argyros: "A. Beat synchronous dance animation based on visual analysis of human motion and audio analysis of music tempo," In *Proc. ISVC* 2013 (2013), pp. 118-127.

[14] T. Shiratori, A. Nakazawa and K. Ikeuchi: "Synthesizing dance performance using musical and motion features," In *Proc. ICRA* 2006 (2006), pp. 3654-3659.

[15] P. Sousa, J. L. Oliveira, L. P. Reis and F. Gouyon: "Humanized robot dancing: Humanoid motion retargeting based in a metrical representation of human dance styles," In *Proc. EPIA* 2011 (2011), pp. 392-406.

[16] T. Takeda, Y. Hirata and K. Kosuge: "Dance step estimation method based on HMM for dance partner robot," *IEEE Transactions on Industrial Electronics* Vol. 54, No. 2 (2007).

[17] T. Taniguchi, K. Hamahata and N. Iwahashi: "Unsupervised segmentation of human motion data using sticky HDP-HMM and MDL-based chunking method for imitation learning," *Advanced Robotics* Vol. 25, No. 17 (2011).

[18] J. M. Wang, D. J. Fleet and A. Hertzmann: "Gaussian process dynamical models for human motion," *IEEE Transactions on Pattern Recognition and Machine Intelligence* Vol. 30, No. 2 (2008).

[19] 後藤, 吉井, 藤原, M. Mauch, 中野: "Songle : 音楽音響信号理解技術とユーザによる謝り訂正に基づく能動的音楽鑑賞サービス," 情報処理学会論文誌, Vol. 54, No. 4 (2013), pp. 1363–1372.