論　文

歌唱時の身振りと音楽との関連性に基づく音楽からの動作生成

青塚　寛之[†]　　山本　正信[††a)]

Dance from Music Based on Gesture in Singing

Hiroyuki AOZUKA[†] and Masanobu YAMAMOTO[††a)]

あらまし　本論文では，歌唱時の身振りと音楽との関連性を教示し，その教示に基づき音楽から動作を生成する手法を提案する．歌唱時のビデオ映像からモーションキャプチャにより動作を測定する．この一連の動作データを動作と音のビートが一致する時点で基本動作列に分節化する．基本動作と同期している音楽フレーズを対応づけ，データベースに登録しておく．動作の生成では，与えられた音楽に最も似た音楽フレーズをデータベースから選択し，この音楽フレーズに対応づけられた基本動作を取り出す．生成された動作は動作主の特徴を継承していることが確かめられた．

キーワード　動作の特徴，パフォーマンスアニメーション，音楽

1. まえがき

会話や音楽にマッチした動作の作成は，アニメーションの分野では古くから行われている．作られた動作のもっともらしさは，アニメータの技量によることが多かったが，最近では音声学や言語学の知識を利用しコンピュータによる作成も試みられている．一方，モーションキャプチャを使えばリアルな動作が容易に再現できる．しかし，測定された動作に限りがあるため，状況に応じて動作を改編させる手法[17]が提案されている．ただし，動作の改編はもとの動作の微調整であり，新たに動作を作り出すのは難しい．そこで，キャプチャされた動作を基本動作に分節化し，基本動作の組合せにより新たな動作を構成するという試み[1],[9]が盛んになってきている．

本論文では，歌唱時の動作と音楽との関連性を教示し，その教示に基づき音楽から動作を生成する手法を提案する．このとき，生成された動作はもとの動作の特徴を継承しているのだろうか．本論文では，分節化と合成により作られた新たな動作にももとの動作の特徴が継承可能であることを示す．

† 新潟大学大学院自然科学研究科，新潟市
Graduate School of Science and Technology, Niigata University, 2–8050 Ikarashi, Niigata-shi, 950–2181 Japan
†† 新潟大学工学部情報工学科，新潟市
Faculty of Engineering, Niigata University, 2–8050 Ikarashi, Niigata-shi, 950–2181 Japan
a) E-mail: yamamoto@ie.niigata-u.ac.jp

2. 関連研究

会話からのしぐさの作成では，口の動きや顔の表情を作り出すことが課題である．実際の会話映像から音素（phoneme）と画像の関連性を学習し，この関連性を利用して音声から会話の映像を生成する手法が提案されている[3],[4]．全身の動作も生成可能ではある[2]が，音素のような短い時間での音と画像の関連付けは，慣性のある身体の動作生成には限界がある．

Stoneら[15]は音素より長い句（phrase）に対応する動作を基本動作としている．会話の映像から基本動作を手動で切り出し，句と基本動作の関連性を学習しておく．一連の会話文に対し，生成動作が滑らかに進行するように基本動作を選択している．

会話に対し音楽からの動作生成の研究例は少ない．Kimら[7]は，身体の関節角加速度のゼロクロスを動作のビートとし，主要なビート間隔で動作を等分に分節化している．得られた基本動作間の遷移確率をあらかじめ学習しておき，動作の生成では，遷移確率に基づき基本動作を逐次加えることにより連続動作を得ている．Kimら[7]は社交ダンスの動作生成を志向しており，そこでは，定まった振付けを一定のリズムで演じることによって動作が構成される．

音楽の主要な特徴はリズムと音の強度変化である．Shiratoriら[13],[14]は，動作データからも動作のリ

ズムと強度変化を抽出した．そして，与えられた音楽のリズムと強度変化に同調した動作セグメントを取り出し，得られた動作セグメント列を連結し音楽に同調した舞踏動作を生成している．これは，舞踏動作のリズムと盛り上がりは音楽のリズムと盛り上がりに同調するという仮定に基づくものである．

Kim ら [7] や Shiratori ら [14] の研究では，音楽のビートやリズムや音の強度に合わせたダンスの生成を志向しているが，個性的な動作の特徴まで考慮したものではない．

これに対し，歌唱時には同じ楽曲でも歌い手により振りが異なってくる．これまで，コンピュータで作成された動作の個性については，ほとんど議論されてこなかった．アニメーションキャラクタの質を向上させるためには，動作の個性にも着目すべきである．

本論文では，音楽フレーズと基本動作を歌手ごとに対応づけ，データベースに登録しておく．動作の生成では，与えられた音楽に最も似た音楽フレーズをデータベースから選択し，この音楽フレーズに対応づけられた基本動作を取り出す．得られた基本動作列を連結することにより，歌い手の個性を反映させた動作を生成することができる．

本論文では音楽フレーズと基本動作を対応づけているが，Hsu ら [6] はアクションとリアクションの対応付けに着目した．彼らはこの対応付けを利用して，社交ダンスのリーダの踊りからパートナーの踊りを推定している．

また，Lee ら [10] は BGM がアニメーションに同調するように，音楽とアニメーション動作の双方を調整する手法を提案している．この手法は微調整に限られるが，音楽に同調した動作の質を向上させるためには有効な手法である．しかし，今回はこの手法を使用していない．

次章では，動作の分節化と音楽との関連付けについて，**4.** では関連付けを利用して音楽から動作を生成する手法について述べる．**5.** では，生成された動作がもとの動作主の個性を継承していること確認する．

3. 音と動作の関連付け

歌唱時の身振りと音を関連づける方法を述べる．音はビートに着目して分節化する．動作は加速度に着目して分節化する．音と動作の分節時刻のうち，ほぼ同時に起こる分節時刻を抽出する．抽出された分節時刻で音と動作を改めて分節化し，音楽フレーズと動作セグメントを得る．この動作セグメントを基本動作とする．動作セグメントに同期した音楽フレーズから音の特徴を求める．この音の特徴と基本動作との対をデータベースに登録する．

3.1 音楽・動作データ収集システム

歌唱時の動作と音楽のデータを収集するシステムは，カラオケ（SD カラオケマイク，SY-MK100-K，Panasonic），DV カメラ，パソコン及びモーションキャプチャソフト（MY-Motion Ver.2.0，（株）Cube Inn）から構成される．カラオケからの伴奏に伴って歌唱している身振りを DV カメラで撮影する．映像は 1 秒間に 30 フレームのサンプリング間隔でパソコンに入力される．

身体を図 1 に示す多関節モデルで表す．腰（Waist）を多関節構造の頂点部位とし，他の部位は関節で親部位と連結されている．各部位には固有の座標系が固定され，部位の運動は部位座標系の親部位座標系に関する運動で表される．腰以外の部位はその親部位に関節で固定されているので，運動の自由度は回転運動の 3 自由度である．腰の動きはシーン座標系に関して，並進と回転の 6 自由度で表される．

パソコンに入力された映像から，モーションキャプチャにより身体姿勢の時系列データが得られる．姿勢の測定原理は文献 [19] に示される．連続するフレーム間映像から直接測定されるのは，各部位の角速度と腰の並進速度である．このとき，回転運動は動座標系で表され角速度も小さいとすれば，角速度をオイラー角表現で表したとしても特異点に陥る心配はない．身体姿勢は部位の角速度と腰の並進速度を初期姿勢と初期

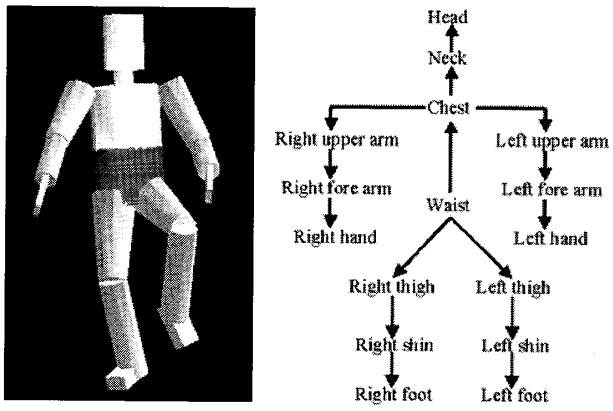

図 1 身体の多関節モデル（左），部位間の接続関係（右）
Fig. 1 Left: An articulated model of human body, Right: A tree of the relations between the body parts.

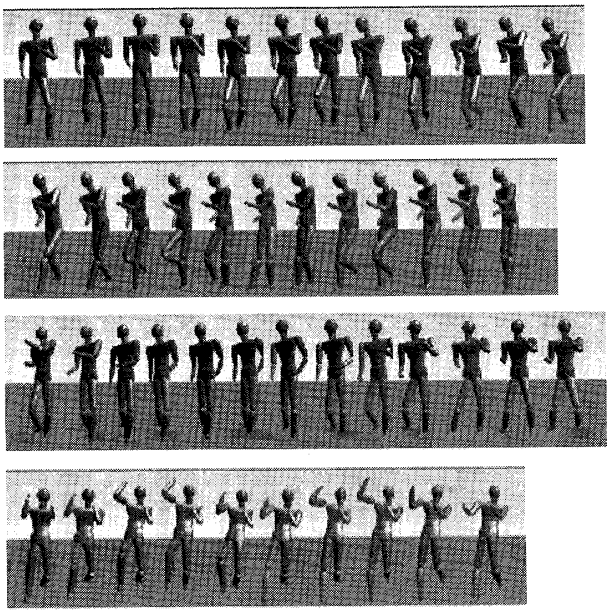

図 2　モーションキャプチャで測定された歌唱身振り
Fig. 2　Captured gestures in singing.

位置に累積することによって得られる．図 2 は，歌唱動作をキャプチャした結果を CG で表示している．全 675 フレームのうち代表的な身振りを四つばかり取り出し 3 フレームおきに表示した．

一方，音楽は DV 映像から分離することによっても得られるが，録音時のノイズが大きいため，カラオケの音源から直接パソコンに入力させる．音は標準では 44.1 kHz でサンプリングされているが，処理の簡便化のため 8 kHz にダウンサンプリングしている．音楽は音声を含まず，伴奏の音圧の時系列データとして録音される．

3.2　ビート抽出

音楽に合わせて体を動かすとき，人間はビートの強弱をとらえていると考えられる．したがって，音楽に合った動作を生成するためには，音楽からビートを抽出することが重要である．ビートはベースやドラムによって刻まれることが多い．したがって，音を周波数成分に分解したとき，低周波成分にビートが表れる．

音のパワースペクトルを $p_i(f)$ とする．ここで i はフレーム時刻，f は周波数である．また，FFT の窓区間を 64 ms とした．次に特定の周波数帯の音の立上り時間を検出するために，周波数 f のパワースペクトルの増分 $d_i(f) = p_i(f) - p_{i-1}(f)$ を算出する．本手法ではビートの周波数帯を低周波の 30〜120 Hz とし，この周波数帯での $d_i(f)$ の総和 D_i を式 (1) によって求める．

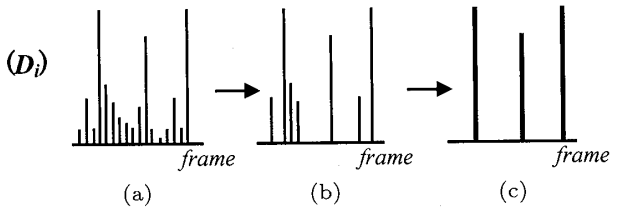

図 3　ビートの検出
Fig. 3　Beats detection.

$$D_i = \sum_{f=f_1}^{f_2} d_i(f) \tag{1}$$

ここで $f_1 = 30\,\mathrm{Hz}$，$f_2 = 120\,\mathrm{Hz}$ である．

音が立ち上がるときは，その音の周波数帯のパワースペクトルの増分 D_i が大きくなる．したがって，あるしきい値以上の D_i を候補ビートとして残す．残ったビートの中で，連続する 3 フレーム中に二つ以上ビート候補がある場合，D_i が最大のビート候補以外は消去する．図 3 はビートの抽出過程を示している．(a) は D_i，(b) はしきい値以上のビート候補，(c) は最終的に得られたビートである．

3.3　動作の分節化

身体の姿勢は関節周りのトルクにより変化し，トルクは筋肉の収縮により発生した力から得られる．身体は，筋肉の収縮と弛緩を繰り返すことにより複雑な動作を作り出していく．本論文では，筋肉が収縮し持続的に力を出している区間の動作を基本動作と呼ぶ．複雑な動作もこの基本動作の組合せから構成される．オイラーの運動方程式によれば，トルクは慣性モーメントと関節角加速度の積に等しい．したがって，身体の力が持続している区間は，関節角加速度により知ることができる．文献 [12] では，茶道点前中の基本動作を関節角加速度により検出している．

図 1 に示すように，身体は腕や脚などの部位から構成され，各部位は固有の座標系をもっている．部位 k の時刻 t における運動は親部位の座標系を基準に，x, y, z 軸周りの角速度で $v_{kx}(t)$, $v_{ky}(t)$, $v_{kz}(t)$ と表す．まず，各部位について角速度から角加速度の大きさの 2 乗 $a_k(t)$ を求める．

$$a_k(t) = \left(\frac{\partial v_{kx}(t)}{\partial t}\right)^2 + \left(\frac{\partial v_{ky}(t)}{\partial t}\right)^2 + \left(\frac{\partial v_{kz}(t)}{\partial t}\right)^2$$

平均角加速度 $\alpha(t)$ を次式で求める．

$$\alpha(t) = \sqrt{\sum_{k=1}^{n} w_k a_k(t)} \tag{2}$$

ここで，$w_k(\geq 0)$ を重み，部位数を n とする．

動作によっては必ずしも全部位が動いているわけではない．平均角加速度の計算を実際に動いている部位（例えば上半身，下半身等）に限定する場合には，限定する部位の重みを 1，限定されていない部位の重みを 0 にする [11]．式 (2) は角速度の 1 階微分を行っている．微分は 1 階でもノイズを伴うため，求めた $\alpha(t)$ は，カルマンフィルタによる平滑化を行っている．

図 4 は **3.2** の手法で音楽から抽出したビートと，$\alpha(t)$ との時間的関係を示しており，上段がビート，下段が平均角加速度 $\alpha(t)$ である．角加速度が極小となる時刻で動作を分割したとき，得られた基本動作は必ずしもビートの間隔と一致しない．動作と音を関連づけるためには，基本動作とビートの間隔は一致していることが望ましい．したがって，ビートを利用して動作の再分節化を行う．

動作の平均角加速度が極小となりかつビートが打たれる時刻を改めて分節時刻とする．実際は，動作の極小時刻とビート時刻との間隔が許容範囲以内であれば，同時に発生したとみなす．図 4 では，右端のビートのみが動作の極小時刻と一致せず，したがって，左の三つのビート時刻で動作と音を分節化した．分節化された個々の動作と音を，それぞれ動作セグメント，音楽フレーズと呼ぶ．

表 1 には，図 2 に示す歌手の動作を分節化した例を示す．下半身（脚）の動きに着目し，歌唱動作を動作と音楽から逐次分節化し，順に動作セグメント番号が割り振られている．また，動作セグメントの長さも示されている．全長 675 フレーム（約 23 秒）の動作が 33 個の動作セグメントに分節化された．

3.4 音楽フレーズの特徴抽出

後節では音楽から動作の生成のために，音楽フレーズ同士の照合を行う．このとき，生の音圧データの照合よりも音の特徴による照合の方がノイズに強い結果が得られる．音声なら音声認識の特徴 [16] を，音楽なら音階やリズムなどの特徴 [5] を抽出する手法がある．本研究の音信号にはカラオケの伴奏音であり音声や環境ノイズの混入はないが，様々な音源が混在しており音楽の特徴を正しく抽出するのは容易ではない．

一方，カラオケはその音源が MIDI 規格で記述されているので，音信号を介さなくても MIDI 記述から音楽の特徴を詳しく知ることが可能である．しかし，この MIDI 記述は現在未公開であるので，処理の簡便さを第一に，音信号を自己回帰移動平均モデル（Auto-Regressive Moving Average model，ARMA モデル）でモデル化する．モデルを音信号に当てはめることにより，音の特徴をモデルのパラメータ値として推定する．

ARMA モデルは次式で表される．

$$y_n = \sum_{j=1}^{m} a_j y_{n-j} + v_n \qquad (3)$$

ここで，y_n と y_{n-j} は，時系列の現在と過去の観測値であり，音圧の時系列である．m と a_j はそれぞれ，自己回帰の次数，自己回帰係数である．次数 m は 3 とした．これは，音楽フレーズ間のパラメータ照合で安定した結果が得られるような，最小セットのパラメータ数として実験的に決定した．また，v_n は平均 0，分散 σ^2 の正規分布に従う白色雑音である．推定すべきパラメータは，係数の a_j と σ^2 である．

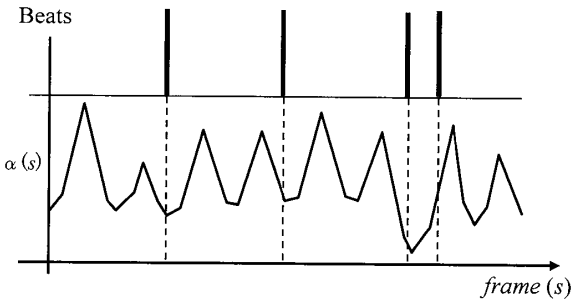

図 4 ビートと角加速度の関係
Fig. 4 Relation of beats and angular accelerations.

表 1 歌手の動作の分節化例
Table 1 Example of segmantation of action in singing.

Action seg. #	Span (*frame*)	Time (*frame*)	Action seg. #	Span (*frame*)	Time (*frame*)
A_1	34	34	A_{18}	11	318
A_2	33	67	A_{19}	34	352
A_3	9	76	A_{20}	24	376
A_4	13	89	A_{21}	70	446
A_5	23	112	A_{22}	13	459
A_6	6	118	A_{23}	32	491
A_7	5	123	A_{24}	11	502
A_8	59	182	A_{25}	45	547
A_9	9	191	A_{26}	11	558
A_{10}	9	200	A_{27}	11	569
A_{11}	25	225	A_{28}	8	577
A_{12}	6	231	A_{29}	5	582
A_{13}	7	238	A_{30}	9	591
A_{14}	11	249	A_{31}	26	617
A_{15}	41	290	A_{32}	12	629
A_{16}	8	298	A_{33}	46	675
A_{17}	9	307			

本手法では最ゆう法[8]により，ARMAモデルの最適なパラメータを推定する．これは，音圧をARMAモデルに当てはめたときの対数ゆう度を計算し，それが最大になるようなパラメータを推定する．パラメータを$\theta=(\sigma^2,a_1,\cdots,a_m)$としたとき，対数ゆう度$l(\theta)$は式(4)によって表される．

$$l(\theta)=-\frac{N}{2}\log 2\pi-\frac{1}{2}\sum_{n=1}^{N}\log d_{n|n-1}-\frac{1}{2}\sum_{n=1}^{N}\frac{(y_n-y_{n|n-1})^2}{d_{n|n-1}} \quad (4)$$

ここでNはデータ数，$y_{n|n-1}$と$d_{n|n-1}$は過去の値から求められる推定値であり，カルマンフィルタの計算により求められる[8]．本手法では，式(4)を最適化するために準Newton法を用いている．

以上により求められたパラメータのうち，σ^2を除いたパラメータを音パラメータとし，音楽フレーズがもつ音の特徴とする．そして音楽フレーズと同期している動作セグメントに対応づけ，データベースに以下の形式で登録する．

{動作セグメント,音パラメータ,区間幅}

将来，MIDI記述が公開されれば，より詳しく音の特徴を知ることができる．ARMAパラメータを含めた音の特徴の中で，有効な特徴を探る試みは今後の課題である．

4. 音楽からの動作生成

音楽が与えられたとき，教示動作に基づき，音楽に対応した動作を生成する手法について述べる．まず，入力音圧からビートを求める．ビート時刻で区切られた区間をビート区間と呼ぶ．ビート区間の音に対して，**3.4**で示した手法により音パラメータを求め，音の特徴とする．各ビート区間ごとに教示動作データの中から，長さと音の特徴が最も良くマッチする動作セグメントを取り出す．ただし，与えられた音楽のビート間隔は，教示時の音楽のビート間隔と同じとは限らない．そのため，ビート間隔と同じ長さの動作セグメントが存在しない場合もある．そこで，まずビート区間の長さを動作セグメントの区間にマッチするように修正する．ついで，新たなビート区間に長さがマッチする動作セグメントの中から，音の特徴が最も良くマッチする動作セグメントを選ぶ．得られた動作セグメント列を入力音楽に最もふさわしい動作系列とする．この動作系列は，必ずしももとの動作の順序どおりではない．そのために生じる姿勢の不連続点は，補間によって滑らかに接続させる．最後に，生成動作をCGモデルで再生し，入力音楽と合成することによって，生成映像を作成する．

4.1 ビート区間幅の修正

与えられた音楽をビートで区切りビート区間列が与えられているとする．教示動作セグメントも与えられたとき，各ビート区間を教示動作セグメントにマッチする長さに次の手順で修正する．修正された区間列を蓄える出力区間列を用意し，最初は空にしておく．

（1） ビート区間列からビート区間を順番に一つ取り出す．取り出すビート区間がなければ処理を終了．

（2） ビート区間の長さが，最も長い動作セグメントを超えれば（6）へ進む．そうでなければ（3）へ進む．

（3） ビート区間と同じ長さをもつ動作セグメントが存在すれば，そのビート区間は修正せずに出力区間列に加える．そして（1）に進む．そうでなければ（4）へ進む．

（4） ビート区間に最も長さの近い動作セグメントを探し，長さの差を許容幅とし，ビート区間と許容幅の対を記憶する．

（5） ビート区間をその次のビート区間と合体させ，新たなビート区間とし，（3）に進む．

（6） （4）で記憶したビート区間と許容幅の対のうち，最も小さな許容幅をもつビート区間を取り出す．

（7） 取り出されたビート区間が複数あるとき，最も短いビート区間を選択しそれを出力区間列に加え，（1）に進む．

新たなビート区間のうち許容幅付のビート区間は，ビート区間幅 ± 許容幅の長さの動作セグメントに対応していることになる．許容幅の付いていないビート区間にはそれと同じ長さの動作セグメントが対応している．

表1の教示用の歌唱動作を使って，与えられた音楽のビート区間を修正した例を示す．まず，この音楽をビートで分節した結果，63個のビート区間が得られた．表2の第1，7列は与えられた音楽のビート区間，第2，8列はその長さを示している．このビート区間を，表1の教示動作セグメントにマッチするように上記の手順で修正する．ビート区間は動作セグメントの短い順に比較する．短い順に並べた教示動作セグメン

表 2 音と動作の関連性に基づく動作の生成例
Table 2 An example of synthesized action.

Interval	Span	Interval	Span	Error	Action	Interval	Span	Interval	Span	Error	Action
T_1	9	I_1	9	0	A_{10}	T_{33}	7				
T_2	7	I_2	7	0	A_{13}	T_{34}	6				
T_3	14	I_3	41	0	A_{15}	T_{35}	8	I_{23}	8	0	A_{16}
T_4	13					T_{36}	13	I_{24}	13	0	A_{22}
T_5	14					T_{37}	20	I_{25}	58	1	A_8
T_6	7	I_4	7	0	A_{13}	T_{38}	9				
T_7	8	I_5	8	0	A_{16}	T_{39}	14				
T_8	13	I_6	13	0	A_4	T_{40}	7				
T_9	6	I_7	6	0	A_{12}	T_{41}	8				
T_{10}	8	I_8	8	0	A_{28}	T_{42}	6	I_{26}	6	0	A_6
T_{11}	8	I_9	8	0	A_{16}	T_{43}	7	I_{27}	7	0	A_{13}
T_{12}	7	I_{10}	7	0	A_{13}	T_{44}	8	I_{28}	8	0	A_{16}
T_{13}	14	I_{11}	25	0	A_{11}	T_{45}	6	I_{29}	6	0	A_{12}
T_{14}	11					T_{46}	29	I_{30}	29	3	A_{31}
T_{15}	10	I_{12}	25	0	A_{11}	T_{47}	13	I_{31}	13	0	A_{22}
T_{16}	8					T_{48}	14	I_{32}	14	1	A_{22}
T_{17}	7					T_{49}	15	I_{33}	15	2	A_{22}
T_{18}	6	I_{13}	6	0	A_{12}	T_{50}	13	I_{34}	13	0	A_{22}
T_{19}	8	I_{14}	8	0	A_{16}	T_{51}	20	I_{35}	43	2	A_{25}
T_{20}	5	I_{15}	5	0	A_7	T_{52}	9				
T_{21}	16	I_{16}	58	1	A_8	T_{53}	14				
T_{22}	13					T_{54}	13	I_{36}	13	0	A_{22}
T_{23}	14					T_{55}	15	I_{37}	58	1	A_8
T_{24}	7					T_{56}	14				
T_{25}	8					T_{57}	7				
T_{26}	13	I_{17}	13	0	A_{22}	T_{58}	22				
T_{27}	14	I_{18}	14	1	A_4	T_{59}	13	I_{38}	13	0	A_{22}
T_{28}	8	I_{19}	8	0	A_{28}	T_{60}	8	I_{39}	8	0	A_{16}
T_{29}	7	I_{20}	7	0	A_{13}	T_{61}	11	I_{40}	11	0	A_{18}
T_{30}	14	I_{21}	25	0	A_{11}	T_{62}	10	I_{41}	10	1	
T_{31}	11	I_{17}				T_{63}	8	I_{42}	8	0	A_{16}
T_{32}	10	I_{22}	23	0	A_5						

トの長さを次に示す．最長の動作セグメント長は 70 であった．

$$\{5,6,7,8,9,11,12,13,23,24,25,26,32,$$
$$33,34,41,45,46,59,70\} \quad (5)$$

まず，表 2 のビート区間 T_1 と T_2 は，集合 (5) の中に長さ 9，7 の動作セグメントが存在するので，区間の修正はしない．そのまま新たなビート区間 I_1，I_2 とする．ビート区間 T_3 には，長さが一致するセグメントはない．T_3 と次のビート区間 T_4 を合体させ，長さ 27 のビート区間 $T_{3,4}$ を作る．このビート区間にも長さが一致するセグメントはない．ビート区間 $T_{3,4}$ を次のビート区間 T_5 と合体させ，長さ 41 のビート区間 $T_{3,4,5}$ を作る．このビート区間の長さは集合 (5) の中に存在する．したがって，$T_{3,4,5}$ を新たなビート区間 I_3 とする．

次に，ビート区間がいずれの動作セグメントとも長さが一致しなかった例を示す．ビート区間 T_{26} は同じ長さの動作セグメントが存在するので，そのままビート区間 I_{17} とする．次の長さ 14 のビート区間 T_{27} には，同じ長さの動作セグメントがない．この区間と最も長さの近いセグメントとの差 1 を許容幅とし，対 $(T_{27},1)$ を作り記憶する．T_{27} と次のビート区間 T_{28} を合体させ，長さ 22 の区間 $T_{27,28}$ を作る．この新しい区間にも長さの一致する動作セグメントはない．この区間と最も長さの近いセグメントとの差 1 を許容幅とし，対 $(T_{27,28},1)$ を作り記憶する．更に，$T_{27,28}$ と次の区間 T_{29} を合体させ，長さ 29 の区間 $T_{27,28,29}$ を作る．このビート区間にも長さの一致する動作セグメントはなく，対 $(T_{27,28,29},3)$ を作り記憶する．この後ビート区間を次々に合体させるが，最長セグメント長 70 を超えるまでに長さの一致する動作セグメントはない．許容幅との対，$(T_{27,28,29,30},2)$，$(T_{27,28,29,30,31},5)$，$(T_{27,28,29,30,31,32},5)$ を記憶する．記憶したビート区間と許容幅の対のうち，最も許容幅の小さな対は $(T_{27},1)$ と $(T_{27,28},1)$ である．この二つの区間のうち短い区間 T_{27} を選び新たなビート区間 I_{18} とする．

表 2 の第 3，9 列は修正手続きで得られた新たなビート区間が示されている．第 4，10 列は区間幅（フレーム数）であり，第 5，11 列は許容幅（フレーム数）である．

4.2 動作セグメントの選択

新たに決定したビート区間にふさわしい動作セグメントを選択する．まず，ビート区間に対し，**3.4** の手法により，音パラメータ θ を求める．次に，ビート区間（またはビート区間 ± 許容幅）と同じ長さの動作セグメントを取り出す．動作セグメントに対応づけられている音パラメータと，ビート区間の音パラメータを比較する．音パラメータ間のユークリッド距離が最小となる動作セグメントを，ビート区間に対応した動作セグメントとし，動作系列を決定する

表 2 の第 6，12 列には，各ビート区間に対応した動作セグメントが示されている．例えば，最初の長さ 9 フレームのビート区間 I_1 は，表 1 の A_3，A_9，A_{17}，A_{30} の動作セグメントが対応可能である．そのうち，ビート区間と音パラメータが最も近かったのは動作セグメント A_9 であった．

一方，許容幅付きのビート区間の例を示す．長さ 14 のビート区間 I_{18} には許容幅 1 がある．長さ 13 あるいは 15 の動作セグメントは，A_4 と A_{22} である．このうちビート区間と音パラメータが最も近かったのは

動作セグメント A_4 であった．

4.3 不連続点の平滑化

決定された動作セグメント列を順につないで動作を生成する．このとき，動作が滑らかにつながらない原因が二つある．一つは，もとの動作で連続していない動作セグメントをつないだとき，つなぎ目で起こる姿勢角と位置のギャップである．もう一つは，ビート間隔が動作セグメントの長さと一致していない場合に起こる時間のギャップである．それぞれつなぎ目を補間し動きの平滑化を行う．

まず．時間ギャップは実験ではたかだか数フレーム（表 2 では 3 フレーム）であったので，動作セグメントを伸縮させることによってギャップを埋める．姿勢角の補間法には線形補間 [9] や平準化補間 [1] 等が提案されている．本論文では線形補間を使用した．

動作セグメントには，本来前後の動作が存在した滑らかな動作である．そこで，この前後の動作を利用して，姿勢の滑らかな推移を身体の各部位ごとに作成する．接続時刻前後に $h/2$ フレーム幅の補間区間を設け，動作セグメント 1 から動作セグメント 2 の動作に移行させるとする．動作セグメント 1 の補間開始時の姿勢を直交行列 $\mathbf{R}_1$，動作セグメント 2 の補間終了時の姿勢を $\mathbf{R}_2$ としたとき，姿勢の差 $\mathbf{R}_1^{-1}\mathbf{R}_2$ を軸周りの回転で表す．軸ベクトルを $\mathbf{n}$，回転角を θ とする．この回転角を区間幅 h で按分する．補間開始時刻から l フレーム目の姿勢は，軸 $\mathbf{n}$ 周りに $l \times \theta/h$ 回転させた回転行列を初期姿勢 $\mathbf{R}_1$ に掛けることによって得る．

補間区間 h の長さは長いほど動作が滑らかに推移する．しかし，動作の特徴を保持するためには補間区間は短いほど良い．本研究では，補間区間の長さを最も短い動作セグメント幅の 2 倍とした．表 2 の動作系列を例にとると，長さ 5 の動作セグメント A_7 が最も短いので，$h = 10$ フレームとした．このとき A_7 は 2 回の補間を受けるので，前後の動作セグメントのつなぎの役割でしかなくなる．

位置の補間は，木構造の最上位の部位である腰部（Waist）のみで行う．腰部の並進運動を腰部位置のフレーム間増分で表しておく．動作セグメント列を順につないだとき，腰部の位置は初期位置に並進運動を累積することにより得られる．腰部の姿勢も補間処理により滑らかに推移するので，並進運動も滑らかに推移する．

本来連続していない動作セグメントを連結し，新たな動作を作成したとき，床面上を足がすべるという不自然な現象 [14] が生じることがある．この原因は，足の運びと全身の並進運動との整合性が保たれていないことから来ている．この不自然さは，歩行や舞踏など身体の並進運動が大きな場合に顕著に表れる．本論文で対象としている歌唱時の振りは，足踏みをするような並進移動量が少ない動作である．そのため，足の運びから腰の並進移動量を再計算するなどの補間は行っていない．図 5 に補間結果の例をアニメーションで示す．補間後の滑らかな動作の変化が確認できる．

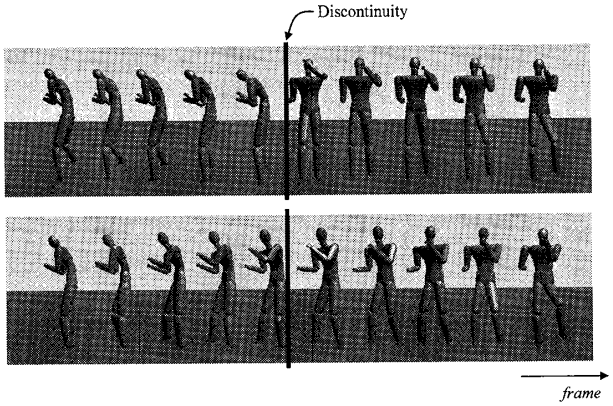

図 5 補間結果の例（上：補間前　下：補間後）
Fig. 5 Example interporation. (Up) Before, (Down) After.

本研究では，約 20 秒間の動作中の 33 個の動作セグメントを使用している．少ない数の動作セグメントでも楽曲にマッチした動作を生成することが可能であるが，動作セグメントの連結部で双方の姿勢が大きく隔たっている場合には，前後で動作の平滑化を行っても不自然な動作になる可能性がある．これを防ぐには，連結させる動作間で動作が滑らかに推移することを拘束条件として，動作セグメントを選択する手法 [7], [14] が提案されている．そのためには，より多くの動作セグメントが必要となる．

5. 動作の特徴の継承

分節化された基本動作を組み合わせて新たに動作を合成したとき，この合成動作はもとの動作の特徴をどの程度継承しているのであろうか．本章では，動作の特徴の継承性を評価する．

身振りに特徴のある歌手について，身振りと音楽との関連性を教示しておく．次に，教示時とは異なる楽曲を与え動作を合成する．一方，この歌手にこの楽曲を歌わせその動作を観測する．合成動作が観測した実動作に似ていれば，動作の特徴は継承されたといえる．

この類似性を人が動作を観察したときの印象により評価する．

合成動作と実動作を比較する映像を次のように作成した．今回の実験では，教示動作の分節化を

- 脚部
- 腕部
- 全身

に着目して行った．つまり着目部位ごとに3種類の動作の分節化を用意した．これは，動作の特徴が身体の部位に偏在している可能性を調べるためである．

身振りに特徴のある3名の歌手A，B，Cを選び，着目部位の分節化ごとに身振りと音楽との関連性を教示した．すなわち，関連性の教示の種類は9種類になる．一方，教示用とは異なる楽曲に対して，それぞれの歌手の合成動作を作成した．次に，動作合成用の楽曲を歌手A，B，Cに歌わせ，実動作をモーションキャプチャで測定した．同一の楽曲に対し，九つの合成動作と三つの実動作が得られたことになる．それぞれの動作から楽曲の伴奏に同期させたアニメーションを作成した．歌手の画像ではなくアニメーションを作成したのは，評価時に歌手の特定による予断を防ぐためである．また，音声を含まない伴奏のみを使用したのも同様の理由である．

教示用楽曲として，「ガッツだぜ!!」（No：TODT-3628），合成用楽曲として，「SHAKE」（No：VIDL-10820）を使用した．いずれも音源のSDカラオケマイク（SY-MK100-K，Panasonic）を利用した．

歌手A，B，Cの合成動作と実動作の映像を被験者に鑑賞させ評価させる．実験は以下の手順で行う．

(1) 歌手A，B，Cの実動作を被験者に鑑賞させ，3名の歌手の特徴を把握させる．図6にディスプレイ上に表示した実動作を示す．

(2) ディスプレイ上に二つの映像（左：合成動作，右：実動作）を提示する．(1)で挙げた歌手の特徴に着目し，合成動作がどの程度実動作と類似しているかを判定させる．評価は5段階評価で，評価の理由も示させる．図7に一対評価時の画面を表示する．

(3) 映像は何回鑑賞してもよく，被験者のペースで実験を進めてもらう．

類似度の5段階評価は

1. かなり似ている　4点
2. 少し似ている　3点
3. どちらともいえない　2点
4. あまり似ていない　1点
5. 全く似ていない　0点

とし，それぞれ上から4〜0点の点数を付ける．被験者は10人であった．最高点が全員1.の評価をしたときの40点であるため，40点満点で点数の総和を求める．

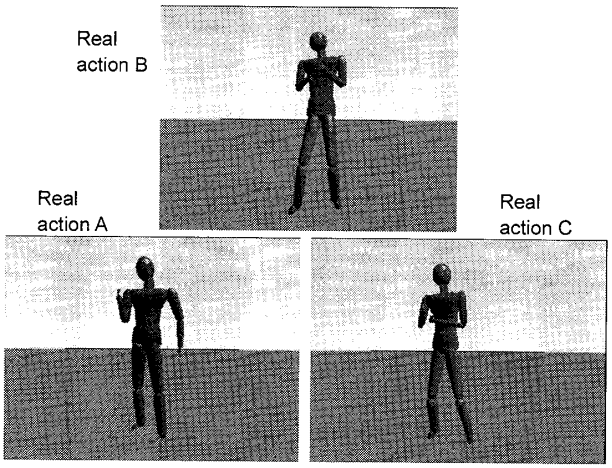

図6 歌手A，B，Cの実動作のアニメーション表示
Fig. 6 Replay actions of singers A, B and C in animation.

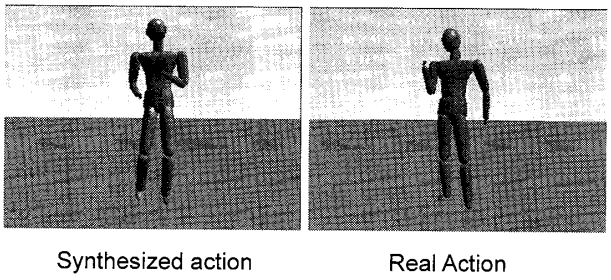

図7 印象による比較実験画面
Fig. 7 Impressive comparison between synthesized and real actions.

1名の被験者が評価にかける時間は20分程度であった．その内訳は，3名の実動作を比較する時間が5分かかる．合成動作の映像は全部で9本あり，1本の長さは20秒かかる．一つの合成動作と3名の実動作を一対ずつ比較するので，27対の映像比較に15分程度必要であった．

図8は評価結果を示したものである．図8の各行は，上から順に脚部，腕部，全身で分節化した場合の結果を示す．

図8の第1行左端のグラフは，左端の棒グラフから順に，歌手A，B，C，の実動作と歌手Aの合成動作との類似度を示している．同図第1行真中のグラフは，歌手Bの合成動作と実動作との比較，同図第1行右端のグラフは，歌手Cの合成動作と実動作の比較を示している．合成動作の手本になる歌手が実動作の歌

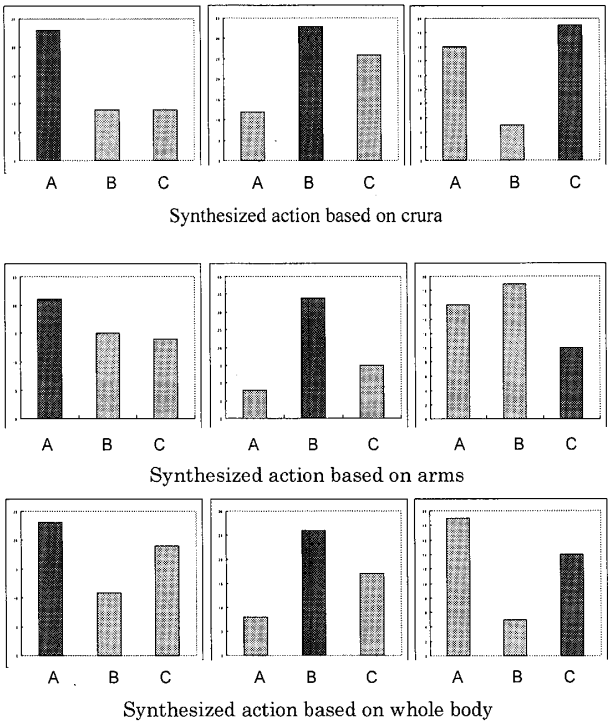

図 8　動作の特徴の継承結果
Fig. 8　Succession of individual features in synthesized actions.

手と同じ場合，棒グラフを濃く描いている．第 1 行の三つのグラフは，いずれも濃い棒グラフが他よりも高い．このことは，どの合成動作も手本とした歌手に他の歌手よりも動作が類似していることを示している．したがって，動作の特徴は継承されたといえる．

一方，第 2 行，第 3 行の右端のグラフでは，濃い棒グラフは最高値ではない．これは，動作の特徴が腕部あるいは全身にあると考えたとき，歌手 C の合成動作が，歌手 C よりも他の歌手に似ていること表している．歌手 C の合成動作が歌手 C の実動作に似ていると答えた人の理由は，脚の動きが似ているというものが最も多かった．これは，第 1 行に示したように，脚の動きに着目した動作の合成結果から理解できる．一方，歌手 C の合成動作が歌手 C の動作に似ていないと答えた理由は，上半身の動きの激しさが異なるという点と，全体的に動作の大きさが異なるという点であった．歌手 C の上半身に，教示時に表れた動作の特徴が，比較時には歌手 C よりも歌手 B に強く現れたものと思われる．歌手は楽曲によって動作の大きさやリズムの取り方を変えることもある．この場合は，本手法では動作の特徴を継承することはできない．

6. むすび

歌唱時の身振りと音楽との関連性に基づき，音楽に合わせて動作を生成する手法を提案した．本手法は，歌手の動作を関節角加速度と音ビートで基本動作に分節化し，基本動作と音楽を関連づけておく．この関連付けを利用して，音楽に同期した動作を生成する．合成した動作は，歌手の動作の特徴をある程度継承していることが印象による評価で確認した．

しかし，動作の特徴の継承が確認されない場合もあった．これは，同じ歌手でも曲によって振り付けを変えて歌う場合があるからである．今回は，一つの曲中の約 20 秒の動作を教示に使用したが，より多くの曲で動作と音楽の関連性を教示すれば，より多彩な特徴を含んだ動作を生成することができよう．

本研究では，合成した動作を人が鑑賞したときの印象で評価したが，評価の普遍性を主張するためには客観的な評価が必要となろう．その際，人が動作を評価したときの着目点が重要となる．着目した動作の特徴を算出できれば，客観的な指標となり得る．これは動作の個人識別や感性情報処理の問題ともいえる．

文　献

[1] O. Arikan and D.A. Forsyth, "Interactive motion generation from examples," ACM Trans. Graph., vol.21, no.3, pp.483–490, 2002.
[2] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," ACM SIGGRAPH'97, pp.353–361, 1997.
[3] M. Brand, "Voice puppetry," ACM SIGGRAPH'99, pp.21–28, 1999.
[4] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," ACM SIGGRAPH'02, pp.388–398, 2002.
[5] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," J. New Music Res., vol.30, no.2, pp.159–171, 2001.
[6] E. Hsu, S. Gentry, and J. Popovic, "Example-based control of human motion," ACM SIGGRAPH/Eurographics Sym. Computer Graphics, pp.69–77, 2004.
[7] T. Kim, S. Park, and S. Shin, "Rhythmic-motion synthesis based on motion-beat analysis," ACM SIGGRAPH'03, pp.392–401, 2003.
[8] 北川源四郎，FORTRAN 77 時系列解析プログラミング，岩波書店，東京，1993.
[9] L. Kovar, M. Gleicher, and F. Pighin, "Motion graphs," ACM Trans. Graph., vol.21, no.3, pp.473–482, 2002.
[10] H.C. Lee and I.K. Lee, "Automatic synchronization

of background music and motion in computer animation," Comput. Graph. Forum, vol.24, no.3, pp.353–361, 2005.
[11] J. Lee, J. Chai, P.S.A. Reitsuma, J.K. Hodgins, and N.S. Pollard, "Interactive control of avatars Animated with human motion data," ACM SIGGRAPH'02, pp.491–500, 2002.
[12] 三富文和，藤原冬樹，山本正信，佐藤泰介，"習慣的な行動の確率文脈自由文法に基づくベイズ識別，" 信学論（D-II），vol.J88-D-II, no.4, pp.716–726, April 2005.
[13] 白鳥貴亮，中澤篤志，池内克史，"モーションキャプチャと音楽情報を用いた舞踊動作解析手法，" 信学論（D-II），vol.J88-D-II, no.8, pp.1662–1671, Aug. 2005.
[14] T. Shiratori, A. Nakazawa, and K. Ikeuchi, "Dance-to-music character animation," Comput. Graph. Forum, vol.25, no.3, pp.449–458, 2006.
[15] M. Stone, P. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, and C. Bregler, "Speaking with hands: Creating animated conversational characters from recording of human performance," ACM SIGGRAP'04, pp.506–513, 2004.
[16] 武田一哉，"音声コーパスと音声情報処理，" 信学技報，PRMU2005-129, 2005.
[17] A. Witkin and Z. Popovic, "Motion warping," ACM SIGGRAPH'95, pp.105–108, 1995.
[18] 山本正信，星　昌人，下山　功，五十嵐達也，"サウンドとモーションの対応付けからのキャラクタの動作生成，" 信学技報，PRMU2001-12, 2001.
[19] 山本正信，"ドリフト修正機能を有する動画像からの身体動作推定法，" 信学論（D-II），vol.J88-D-II, no.7, pp.1153–1165, July 2005.

（平成 18 年 12 月 11 日受付，19 年 5 月 15 日再受付）

青塚　寛之

平 16 新潟大・工・情報卒．平 18 同大大学院修士課程了．同年，（株）山形ケンウッド勤務．現在，無線機の量産へ向けた技術開発に従事．在学中は，音楽分析，コンピュータアニメーションなどの研究に従事．

山本　正信　（正員）

昭 48 九工大・工・制御卒．昭 50 東工大大学院修士課程了．同年，電総研入所．動画像処理，コンピュータビジョン等の研究に従事．平元〜2 カナダ国立研究協議会招聘研究員．昭 62 情処学会研究賞受賞．現在，新潟大学工学部情報工学科教授．工博．情報処理学会，IEEE CS 等各会員．