Vol.2016-MUS-112 No.15
2016/8/1

視聴覚統合ビートトラッキングと リアルタイムコード認識を用いたダンス共演ロボット

大喜多 美里[1,a)]　坂東 宜昭[1,b)]　糸山 克寿[1,c)]　吉井 和佳[1,d)]

概要：本稿の目標は音楽音響信号と共演者（人間）のダンスを表す骨格情報を用いて，ビートやコードといった音楽情報を推定しながら踊る共演ロボットの開発である．実用的なロボットだけでなく人が親しみを感じるようなエンターテイメントロボットを開発することは，将来的に人とロボットが共存するために重要な課題である．提案システムは，視聴覚統合ビートトラッキング部，コード認識部，ロボット動作制御部から成る．ビートトラッキングでは，音響信号を用いた手法が数多く提案されてきたが，テンポ変動や裏拍を多く含む楽曲の場合精度が十分でないという問題から，我々は共演者のダンスを表す骨格情報を用いた視聴覚統合ビートトラッキングを提案してきた．この手法は，一意に定めた音響テンポと視覚テンポ尤度の推定誤りが視聴覚統合に影響するという問題があった．本稿では，音響テンポを尤度として求め，さらに視覚テンポ尤度を平滑化して統合することで精度向上を図る．コード認識では，特徴量の出力確率の計算で混合 von Mises-Fisher 分布が混合ガウス分布に比べて有効であることが知られているため，本稿では混合 von Mises-Fisher 分布を用いてリアルタイムでのコード認識を行う．実験では，提案法による視聴覚統合ビートトラッキングの有効性と，ロボットのシミュレータを用いて提案システムの動作を確認した．

1. はじめに

ダンス共演ロボットとは，音楽と人間の動作を認識しながら人間と共に踊るエンターテインメントロボットである．実用的なロボットだけでなく人が親しみを感じるようなエンターテイメントロボットを開発することは，将来的に人とロボットが共存するために重要な課題である．このようなロボットとして，バイオリンを演奏するロボット [1] やボール上でダンスするチアリーダーロボット [2]，人の演奏に合わせてフルートを演奏するロボット [3] が開発されている．特にダンスは多くの文化圏で親しまれており，言葉の壁が存在しないインタラクションであるため誰でも楽しむことが可能であることから，本稿ではダンス共演ロボットに注目する．

ロボットが人と協調して踊るためには，ビートやコードといった音楽情報をリアルタイムで正確に推定する技術や，ロボットの動作を適応させる技術が必要である．現在までに，社交ダンスの動作をインタラクティブに生成することで人と踊る技術 [4]，高度な制御でヒューマノイドロボットによる自然なダンスを実現する技術 [5], 音楽のビートに合わせて足踏みと歌唱を行うロボット [6] が開発されている．本稿では，音楽音響信号と共演者（人間）のダンスを表す骨格情報を用いて，ビートとコードを推定しながら踊る共演ロボットを提案する．

ビートトラッキングでは，音響信号を用いた手法 [7] [6] や視覚情報を用いた手法 [8] [9] など数多く提案されてきた．また，糸原らはギター演奏者と共演するロボットのために，手の動きと音響信号を用いた視聴覚統合ビートトラッキングを提案した [10]．我々はダンス共演ロボットの音楽理解能力の向上のため，共演者のダンスを表す骨格情報を用いた視聴覚統合ビートトラッキングを提案した [11]．しかし音響信号のテンポを一意に定めた後に視聴覚統合を行うために，音響テンポの推定誤りの影響が大きく，また視覚テンポ尤度の推定誤りによる影響が問題であった．本稿では，音響信号からテンポの尤度を推定し，さらに視覚テンポ尤度を平滑化して統合することで精度向上を図る．

コード認識は一般的に，音響信号からの特徴量抽出と確立モデルによる特徴量分類の 2 段階の処理からなる特徴量として，12 のピッチクラス（C, C#, …, B）のエネルギー分布を表した 12 次元クロマベクトル [12] を用いる．特徴量抽出はビートごとに行い，特徴量分類では出力確率に混合 von Mises-Fisher 分布 [13] を用いる．

提案システムは全て ROS（Robot Operating System）[14]

[1] 京都大学
[a)] ohkita@sap.ist.i.kyoto-u.ac.jp
[b)] yoshiaki@kuis.kyoto-u.ac.jp
[c)] itoyama@kuis.kyoto-u.ac.jp
[d)] yoshii@kuis.kyoto-u.ac.jp

上で構築した．これにより，各モジュール間のデータのやりとりを容易に実現できる．さらに強力な可視化システムが内臓されており，ロボットの動作をシミュレータで確認することが可能である．また，複数モダリティから取得したデータをリアルタイムで再生できるという利点から，視聴覚の情報を同時に取り扱う本システムで採用した．本稿では，シミュレータでの動作確認を行う．

2. ダンス共演ロボット

本稿で提案するダンスロボットは，骨格情報と音響信号からリアルタイムでビートとコードを推定し，それらに基づいて動作制御を行う（図 1）．視聴覚統合ビートトラッキング部，コード認識部，ロボット動作制御部から成る．本稿では，音楽はマイクにより取得した音響信号，ダンスは Kinect やモーションキャプチャで取得した骨格時系列情報で表現する．

視聴覚統合ビートトラッキングでは，複数モダリティを観測とした状態空間モデルを用いて音楽とダンスに含まれる情報を統合する．従って，視聴覚統合ビートトラッキング部で扱う問題を以下のように定める．

入力	音響特徴量：$\{A_1, A_2, \cdots, A_k\}$
	骨格特徴量：$\{S_1, S_2, \cdots, S_k\}$
出力	現在のテンポ：ϕ_k
	ビート時刻：θ_k

k は現在のビート数を示す．音響信号と骨格情報からそれぞれ特徴量抽出を行い，特徴量から ϕ_k, θ_t の確率密度を状態空間モデルを用いて推定する．ϕ_k と θ_k から次のビート時刻 θ_{k+1} を予測し，現在時刻が予測された次のビート時刻を過ぎるごとに推定を行う．

コード認識では，コードは 12 のルート音（C, C#, D, …, B）と 2 つのコードの種類 (major/minor) の組み合わせである 24 クラスとする．現在時刻が予測した次のビート時刻 θ_{k+1} に到達する度に，θ_k〜θ_{k+1} 間のコード C_k を出力する．特徴量は 12 次元クロマベクトルを用いる．特徴量分類では出力確率に混合 von Mises-Fisher 分布 [13] を用いる．

ロボットの動作生成はビートごとに行い，ビート時刻 θ_{k+1} でコード C_k を用いて動作を生成する．本システムは ROS [14] 上で構築されている．

2.1 視聴覚統合ビートトラッキング

ダンス共演では音楽と共演者のダンスという 2 つの情報が存在するため，両者を統合することでビートトラッキングの精度向上を図る．[11] では，音響テンポと視覚テンポ尤度の推定間違いが視聴覚統合に影響していた．本稿では，音響信号のテンポを一意に定めず尤度として求め，視

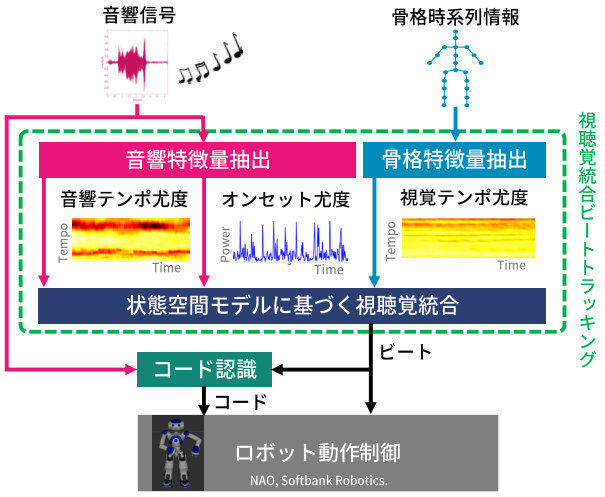


図 1 ダンス共演ロボットシステム概要図

覚テンポ尤度は平滑化させて視聴覚統合することで問題を解決する．音響特徴量抽出には [10] と同様，テンポ変化追従性に優れている村田らの手法 [6] を用いる．骨格特徴量抽出には Chu ら [9] の推定法を応用する．

2.1.1 音響特徴量の抽出

音響特徴量抽出の概要を図 2 に示す．音響特徴量 A_k はオンセット尤度 $F_k(t)$ と音響テンポ尤度 $R_k(u)$ からなる．ここで u はテンポを表す．各時間フレームの音響信号 y_t に対して周波数解析を行い，メル尺度のスペクトログラムを求める．画像のエッジを強調する際に用いられるソーベルフィルタをスペクトログラムを適用することで，パワーが増大している時刻を強調したオンセットベクトル $d(t, f)$ を求める．ここで，f はメルフィルタバンクの次元を表す．オンセット尤度 $F_k(t)$ はオンセットベクトル $d(t, f)$ の周波数成分の要素を足しあわせることで得られる．

$$F_k(t) = \sum_{f=1}^{F_\omega} d(t, f). \tag{1}$$

次に，以下で定義される正規化相互相関マッチングを用いて音響テンポ尤度を求める．

$$R(t, s) = \frac{\sum_{j=1}^{F_\omega} \sum_{i=0}^{P_\omega - 1} d(t-i, j) d(t-s-i, j)}{\sqrt{\sum_{j=1}^{F_\omega} \sum_{i=0}^{P_\omega - 1} d(t-i, j)^2 \sum_{j=1}^{F_\omega} \sum_{i=0}^{P_\omega - 1} d(t-s-i, j)^2}}. \tag{2}$$

P_ω はパターンマッチングの窓幅で s はシフトパラメータである．これにより一般的な自己相関関数を用いるよりも短い窓幅でテンポを抽出する．ここで計算効率化のために Fast Normalized Cross-Correlation [15][*1] を用いた．これにより，$R(t, s)$ が負値を持つため指数関数で変換する．これを $R'(t, s)$ と表す．テンポ u に対応するシフト数を s_u とすると，時刻 θ_k における音響テンポ尤度は $R_k(u) = R'(\theta_k, s_u)$ で得られる．本稿では，[6] と同様に倍テンポの推定誤りを

[*1] http://scikit-image.org/

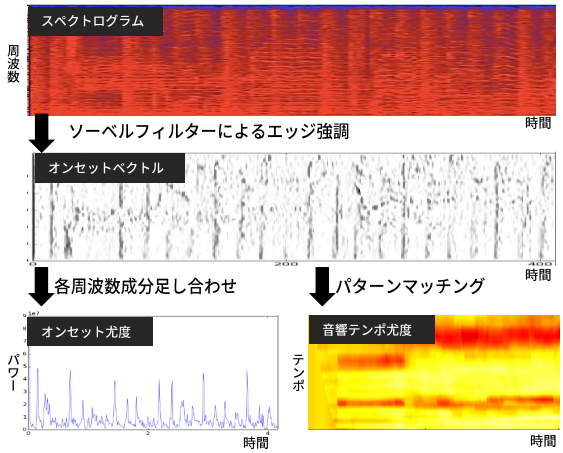


図 2 音響特徴量抽出の概要.

防ぐために，テンポは n beats per minute (BPM) から $2n$ (BPM) に制限している.

2.1.2 骨格特徴量抽出

ビート時刻 θ_k における骨格特徴量 S_k は視覚テンポ尤度 $S_k(u)$ からなる．Chu らはダンス動画に対するオフラインでのテンポ推定手法 [9] を提案した．本稿では [9] をオンライン化し，さらにダンスの細かい動作を考慮できるように骨格情報を用いた手法を提案する.

抽出の概要を図 3 に記す．関節数を J，時刻 t での首や腰などの各関節の 3 次元座標を表す骨格情報を $\{\boldsymbol{b}_1(t),\cdots,\boldsymbol{b}_J(t)\}$ とする．ここでは，ダンサーがビート時刻に関節を停止・回転させる傾向があることに基づき，関節の動作が停止・回転する時刻の周期性に着目することで視覚テンポ尤度を求める.

各関節が停止・回転する時刻の決定 停止時刻は関節の移動距離が極小となる時刻とする．各関節の移動距離を $g_j(i)=||\boldsymbol{b}_j(i+1)-\boldsymbol{b}_j(i)||$ とすると，停止時刻の集合 $\mathcal{I}_j^{\mathrm{st}}$ は以下で得られる.

$$\mathcal{I}_j^{\mathrm{st}}=\left\{\underset{i\le m\le i+n}{\operatorname{argmin}}\ g_j(m)\ \middle|\ t-N+1\le i<t-n\right\},\quad(3)$$

ここで n はシフト長である．また，回転時刻を関節の内積が極大となる時刻とする．内積 $h_j(i)$ を以下で求める.

$$h_j(i)=\boldsymbol{o}_{j,i}^T\boldsymbol{o}_{j,i+1},\quad(4)$$

$$\boldsymbol{o}_{j,i}=\frac{\boldsymbol{b}_j(i+1)-\boldsymbol{b}_j(i)}{g_j(i)}.\quad(5)$$

回転時刻の集合 $\mathcal{I}_j^{\mathrm{tr}}$ は以下で得られる.

$$\mathcal{I}_j^{\mathrm{tr}}=\left\{\underset{i\le m\le i+n}{\operatorname{argmin}}\ h_j(m)\ \middle|\ t-N+1\le i<t-n\right\}.\quad(6)$$

波形の作成 $\mathcal{I}_j^{\mathrm{st}}$ と $\mathcal{I}_j^{\mathrm{tr}}$ は時刻の離散集合であるため，周期性を求めるためにガウス関数を用いて波形に変換する．以下のように，ガウス関数を用いて $\mathcal{I}_j^{\mathrm{st}},\mathcal{I}_j^{\mathrm{tr}}$ から波形 $y_j^{\mathrm{st}}(t),y_j^{\mathrm{tr}}(t)$ を作成する.

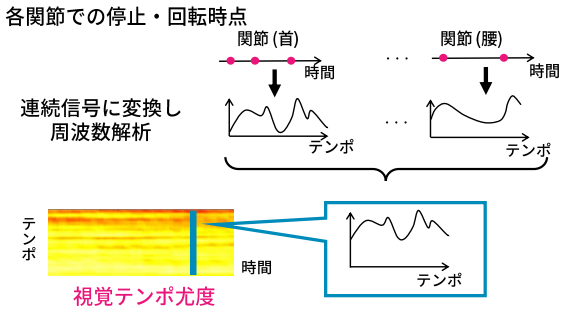


図 3 骨格特徴量抽出の概要

$$y_j^{\mathrm{st}}(t)=\sum_{i\in\mathcal{I}_j^{\mathrm{st}}}\mathcal{N}(t|i,\sigma_y^2),\ y_j^{\mathrm{tr}}(t)=\sum_{i\in\mathcal{I}_j^{\mathrm{tr}}}\mathcal{N}(t|i,\sigma_y^2).\quad(7)$$

$\mathcal{N}(x|\mu,\sigma)$ は変数を x とする平均 μ，分散 σ^2 の正規分布の確率密度関数である.

周波数解析 $y_j^{\mathrm{st}}(t),y_j^{\mathrm{tr}}(t)$ にフーリエ変換を行いスペクトル $\hat{y}_j^{\mathrm{st}}(f),\hat{y}_j^{\mathrm{tr}}(f)$ を求める．それらを全関節で足し合わせたものを $S(t,f)$ とする

$$S(t,f)=\sum_{j=1}^{J}(|\hat{y}_j^{\mathrm{st}}(f)|+|\hat{y}_j^{\mathrm{tr}}(f)|).\quad(8)$$

視覚テンポ尤度 $S_k(u)$ は $S_k(u)=S(\theta_k,f_u)$ で与えられる．ここで f_u はテンポ u に対応する周波数を表す.

2.1.3 状態空間モデルに基づく視聴覚統合

状態空間モデルを用いて，音響特徴量と視覚特徴量の統合を行う（図 4 ）．状態変数 $\boldsymbol{z}_k$ と観測変数 $\boldsymbol{x}_k$ はビート時刻 θ_k，テンポ ϕ_k，音響テンポ尤度 R_k，オンセット尤度 F_k，視覚テンポ尤度 S_k を用いて以下のように表現する.

$$\boldsymbol{z}_k=[\phi_k,\theta_k]^T,\ \boldsymbol{x}_k=[R_k^T,S_k^T,F_k^T]^T.\quad(9)$$

観測モデル 観測変数は全て独立とみなすことで，観測モデルを以下と定義する．ここで，視覚テンポ尤度に ε を加えることで平滑化を行う.

$$\begin{aligned}&p(\boldsymbol{x}_k|\boldsymbol{z}_k)=p(R_k|\boldsymbol{z}_k)p(S_k|\boldsymbol{z}_k)p(F_k|\boldsymbol{z}_k),\quad(10)\\&p(R_k(u=\phi_k)|\boldsymbol{z}_k)\propto R_k(u=\phi_k),\\&p(S_k(u=\phi_k)|\boldsymbol{z}_k)\propto S_k(u=\phi_k)+\varepsilon,\\&p(F_k(t=\theta_k)|\boldsymbol{z}_k)\propto F_k(t=\theta_k).\end{aligned}$$

状態遷移モデル 状態遷移は以下と定義する.

$$p(\boldsymbol{z}_k|\boldsymbol{z}_{k-1})=\mathcal{N}(\phi_k|\phi_{k-1},\sigma_\phi^2)\mathcal{N}(\theta_k|\theta_{k-1}+60/\phi_{k-1},\sigma_\theta^2).\quad(11)$$

推論アルゴリズム 観測変数がガウス分布に従わないため，本状態空間の推定にはパーティクルフィルタを用いる．SIR (Sequential Importance Resamling) パーティクルフィルタ [16] を用いることで，計算を効率化する．提案分布は状態遷移モデルに基づく．ここで，パーティクル

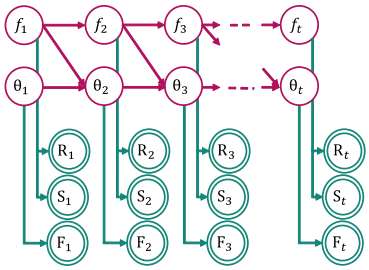


図 4 状態空間モデル.

表 1 比較手法

手法	音響テンポ尤度 (音響特徴量)	視覚テンポ尤度 (骨格特徴量)	オンセット尤度 (音響特徴量)
本手法	✓	✓	✓
視覚テンポ尤度なし	✓		✓
音響テンポ尤度なし		✓	✓

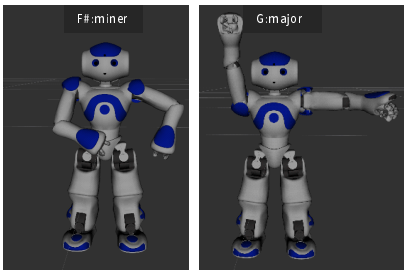


図 5 シミュレータによる動作確認結果.

の過度な集中を防ぎテンポ変化への追従を可能にするため，L' 個のパーティクルをランダムに選び独立して遷移させる．

2.2 リアルタイムコード認識

コードは 12 のルート音（C，C#，D，…，B）と 2 つのコードの種類 (major/minor) の組み合わせである 24 クラスとし，視聴覚統合ビートトラッキング部で予測したビート時刻 θ_{k+1} に到達する毎に θ_k〜θ_{k+1} 間のコード C_k を出力する．コード認識は音響信号からの特徴量抽出と特徴量分類からなる．

特徴量抽出 特徴量として，12 次元クロマベクトル [12] を用いる．各フレームの音響信号に対して高速フーリエ変換でパワースペクトル計算し，時刻 θ_k〜θ_{k+1} 間で足しあわせる．これを線形補間を用いて対数周波数に変換し，クロマベクトルを求める．

特徴量分類 各コード区間における特徴量の出力確率は混合 von Mises-Fisher 分布を用いて求める．クロマベクトルの分布が混合ガウス分布に従うとは限らないことから，混合 von Mises-Fisher 分布の有効性が知られている [13]．コードは 12 のルート音と 2 つのコードの種類の組み合わせである 24 クラスを対象とするが，クロマベクトルの要素を巡回シフトしてルート音を C に統一することで，2 クラス (Cmajor/Cminor) に対してのみ学習する．他のコードのモデルは混合 von Mises-Fisher 分布のパラメータを巡回させることで得られる．パラメータの推定には EM アルゴリズムを用いた．

2.3 ダンス動作生成

NAO のダンスポーズは，24 種類の各コードに対応する 24 種類のダンスポーズを定めた [17]．視聴覚統合ビートトラッキング部で予測したビート時刻 θ_{k+1} 毎に，コード認識部で推定したコード C_k に合わせて動作を生成する．ロボットは SoftBank Robotics の NAO [18] を想定している．NAO のドライバーである NAOqi と ROS [14] を用いてシステムを構築した． これにより，各モジュール間のデータのやりとりを容易に行うことができる．さらに強力な可視化システムが内臓されており，ロボットの動作をシミュレータで確認することが可能である．ロボットの動作速度を考慮し，ビート時刻 θ_k〜θ_{k+1} 間が 0.5ms 以上の場合に動作を生成する．

3. 実験

本手法による視聴覚統合ビートトラッキングの精度評価と，シミュレータを用いた提案システムの動作確認を行った．

3.1 視聴覚統合ビートトラッキング

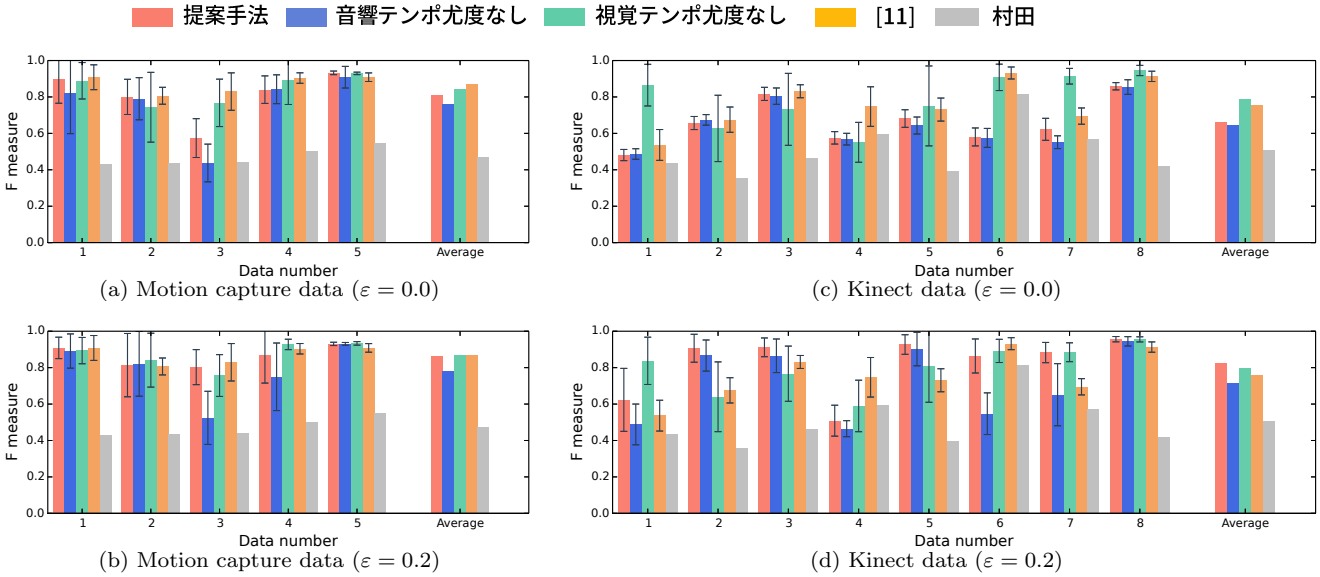
本手法の有効性を確認するため，次の手法と比較する：音響テンポ尤度 $R_k(u)$ の観測を除いた場合，視覚テンポ尤度 $S_k(u)$ の観測を除いた場合，[11]，村田らの手法 [6]．実験には Cyprus[*2] のモーションキャプチャデータ 5 曲と，Kinect で取得したダンス経験者によるポップスの曲に合わせたミックスダンス 8 曲を使用した．後者では，音声や手拍子が含まれている音楽音響信号をマイクで録音した (16kHz, 16bit)．

音響特徴量の計算にはロボット聴覚ソフトウェア HARK [19] を使用した（$n = 90$）．音の立ち上がりのタイミングが 100ms 以上の場合は人間には音がずれて感じられることに基づき [20]，ビート時刻の推定結果と正解結果の差が 100ms 以内ときを正解とした．各データについて適合率 (=推定成功拍数/検出拍総数)・再現率 (=推定成功拍数/正解拍総数) から F 値を求めた．パーティクルの初期値による影響を考慮し，各データに対してパーティクルの初期値をランダムに変更して 30 回推定を行い F 値の平均により評価した．

パーティクル数は $L = 1000$, ε は {0.0,0.2} のそれぞれで評価した．状態遷移におけるパラメータ $\sigma_\phi, \sigma_\theta$ は各手法においてそれぞれ {1.0, 3.0, 5.0}，{0.01, 0.02, 0.03, 0.04} の全組み合わせで実験を行い，全データの平均が最高となるものを選択した．また，[11] でのパラメータ σ_M は {0.25, 4.0, 9.0} から同様に選び，他は [11] と同じである．

図 6 に評価結果を示す．全データにおいて本手法は村田らの手法より F 値が向上している．本手法と観測を制限した 2 つの手法で，$\varepsilon = 0.2$ の場合 $\varepsilon = 0.0$ に比べて両データ

[*2] http://dancedb.cs.ucy.ac.cy/



(a) Motion capture data ($\varepsilon = 0.0$)
(c) Kinect data ($\varepsilon = 0.0$)
(b) Motion capture data ($\varepsilon = 0.2$)
(d) Kinect data ($\varepsilon = 0.2$)

図 6 評価結果. 横軸はデータ番号を表す.

セットの平均が高い．これにより，状態空間モデルにおける骨格特徴量の統合の際に平滑化することが有効であると考えられる．この平滑化によりパーティクルの過度な集中を防ぐことができるため，視覚テンポ尤度に誤りが含まれる場合だけでなく，オンセット尤度の推定誤りによるパーティクルの収束も防ぐことができていると思われる．

以下では $\varepsilon = 0.2$ の場合について述べる．Kinect データでは本手法が最も高い平均 F 値を実現した．モーションキャプチャデータでは，視覚テンポ尤度を用いた場合と [11] と同程度の F 値を示した．Kinect データにおける平均 F 値がモーションキャプチャデータより低いのは，Kinect で取得している関節数がモーションキャプチャより少ないこと，Kinect データに多くのノイズが含まれること，さらに Kinect ではオクルージョンが発生することが原因であると考えられる．

推定結果例を図 7 に示す．図 7-(a)(b) では，視覚テンポ尤度と音響テンポ尤度のピークが正解テンポ付近に存在しており，正解テンポ及び正解ビート時刻に正しく収束している様子が分かる．一方，図 7-(c) では，視覚テンポ尤度が正しく推定できておらず，パーティクルが収束しなかったためビート時刻の推定に失敗している．視覚テンポ尤度の推定が失敗するのは，オクルージョンや手足の動きが細かいダンスの場合に停止・回転時刻が正しく抽出されていないためと思われる．図 7-(d) では，視覚テンポ尤度が正しく推定できていないが，音響テンポ尤度のピークが正解テンポに多く存在するためにパーティクルが正しく収束している．

3.2 シミュレータを用いたシステムの動作確認

提案システムの動作の確認を ROS 上のシミュレータを用いて行った．実験には，4 つのルート音（C,D,E,F）を 2 つのコードの種類 (major/minor) で 80BPM で順に再生するピアノ音源を作成 [*3] して使用した．この音響信号を再生し，ビートに合わせて手を上げ下げする様子を Kinect で取得した．同時に，音響信号はスピーカーの近くに設置したマイクロホンで録音した (16kHz,16bit)．ROS は記録されたデータをリプレイする機能をもつため，採録した状況を ROS 上で再現することができる．今回はこの機能を用いてデータを取得し，シミュレータ実験を行った．

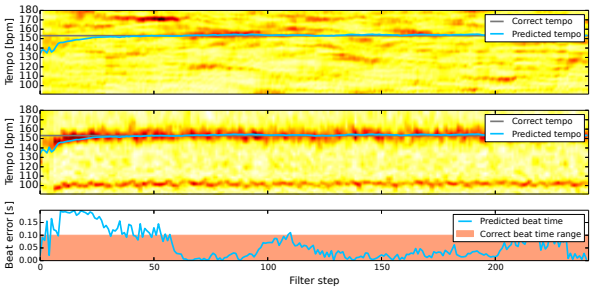
提案システムは，Ubuntu14.04 上で ROSindigo と HARK を使用して実装している．視聴覚統合ビートトラッキング部の BPM は $n = 60$ とする．また，骨格特徴量抽出は 10fps で行った．コード認識における混合 von Mises-Fisher 分布の学習には，The Beatles データセット [*4] の 180 曲の内，12 のルート音と 2 つの和音の種類の組み合わせである 24 クラスに該当するコード区間を使用した．

動作の様子を図 5 に示す．提案システムがリアルタイムで動作することが確認できた．一方，ビート時刻ごとに動作を生成しているため，ダンスポーズとビート時刻に遅延が生じている．また，コード認識ではビート時刻 θ_{k+1} で θ_k〜θ_{k+1} 間のコード C_k を出力しているために，認識したコードは 1 つ前のビート区間のものとなっている．

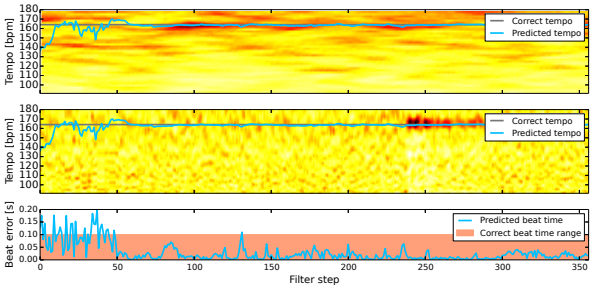
4. おわりに

本稿では，音響信号と共演者の骨格情報からビートとコードを認識して動作するダンス共演ロボットを提案した．実験により，提案法による視聴覚統合ビートトラッキングの有効性と，ロボットのシミュレータによる提案システムの動作を確認した．視聴覚統合ビートトラッキングでは，各特徴量への信頼度を考慮することで精度が向上すると思われる．

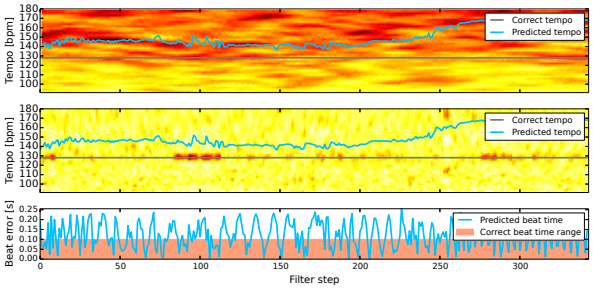
[*3] https://www.noteflight.com/
[*4] http://www.isophonics.net/datasets



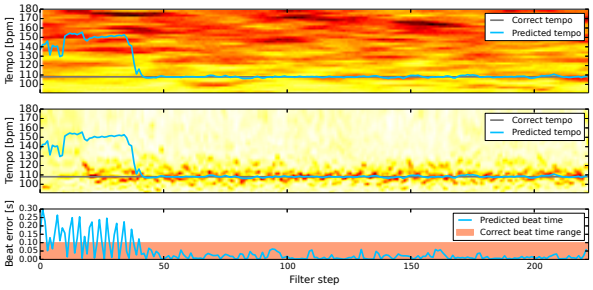
(a) 推定例 (motion capture data No. 1)



(b) 推定例 (Kinect data No. 3)



(c) 推定例 (Kinect data No. 1)



(d) 推定例 (Kinect data No. 7)

図 7 ビートの推定例．上段中段では青線：推定テンポ，灰色線：正解テンポを表す．上段は視覚テンポ尤度，中段は音響テンポ尤度を表す．下段では青線：推定ビート時刻と正解ビート時刻の誤差，赤枠：正解範囲を表す．

混合 von Mises-Fisher 分布を用いたコード認識では，リアルタイムでの動作を確認した．提案システムでは，現在時刻のコードではなく 1 つ前のビート区間のコードを認識している．コードの遷移確率を用いて次のコードを予測することで，現在時刻のコードを予測することができると考えられる．ロボットの動作は，ビート時刻ごとに制御していたため遅延が発生している．遅延時間を考慮して制御することで解決できると考えられる．

謝辞 本研究の一部は，JSPS 科研費 24220006, 26700020, 26280089，16H01744 と JST OngaCREST の支援を受けた．また，コード認識についての丸尾智志氏の多大なるご協力に感謝します．

参考文献

[1] Kusuda, Y.: Toyota's Violin-playing Robot, *Ind. Robot*, Vol. 35, No. 6, pp. 504–506 (2008).
[2] Murata Manufacturing Co., Ltd: Cheerleaders Debut, http://www.murata.co.jp/cheerleaders/ (2015).
[3] Petersen, K. et al.: Development of a Aural Real-Time Rhythmical and Harmonic Tracking to Enable the Musical Interaction with the Waseda Flutist Robot, *IROS* (2009).
[4] Kosuge, K. et al.: Partner Ballroom Dance Robot-PBDR-, *SICE Journal of Control, Measurement, ans System Integration*, Vol. 1, No. 1, pp. 74–80 (2008).
[5] Kaneko, K. et al.: Cybernetic Human HRP-4C, *Humanoids* (2009).
[6] Murata, K. et al.: A Beat-Tracking Robot for Human-Robot Interaction and Its Evaluation, *Humanoids* (2008).
[7] Goto, M.: 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).
[8] Guedes, C. et al.: Extracting Musically-Relevant Rhythmic Information from Dance Movemen by Applying Pitch-Tracking Techniques to a Video Signal, *SMC* (2006).
[9] Chu, W. et al.: Rhythm of Motion Extraction and Rhythm-Based Cross-Media Alignment for Dance Videos, *ACM Multimedia* (2012).
[10] Itohara, T. et al.: Particle-filter Based Audio-visual Beat-tracking for Music Robot Ensemble with Human Guitarist, *IROS* (2011).
[11] Ohkita, M. et al.: Audio-Visual Beat Tracking Based on a State-Space Model for a Music Robot Dancing with Humans, *IROS* (2015).
[12] Fujishima, T.: Realtime chord recognition of musical sound: A system using common lisp music, *ICMC* (1999).
[13] Maruo, S.: Automatic Chord Recognition for Recorde Music based on Beat-Position-Dependent Hidden Semi-Markov Model, Master's thesis, Kyoto (2016).
[14] Quigley, M. et al.: ROS: an open-source Robot Operating System, *ICRA* (2009).
[15] Lewis, J. P.: Fast Normalized Cross-Correlation, *Vision interface* (1995).
[16] Sanjeev, M. et al.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, *Signal Processing* (2002).
[17] Tsumaki, M. et al.: A Humanoid Robot that can Sing and Dance to Music by Recognizing Beats and Chords in Real Time, *ISMIR, Late-Breaking / Demo (LBD)* (2015).
[18] SoftBankRobotics: NAO, https://www.ald.softbankrobotics.com/en/cool-robots/nao (2016).
[19] Nakadai, K. et al.: Design and Implementation of Robot Audition System'HARK'-Open Source Software for Listening to Three Simultaneous Speakers, *Advanced Robotics* (2010).
[20] Rasch, R. A.: Synchronization in Performed Ensemble Music, *J Acta Acustica united with Acustica* (1979).