社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B102-10

マルチロボットによるKinectを用いた同期合奏

Multi-robot synchronized ensemble with Kinect

糸原達彦 † **Tatsuhiko Itohara**
水本武志 † **Takeshi Mizumoto**
Angelica Lim† **Angelica Lim**
大塚琢馬 † **Takuma Otsuka**
中村圭佑 ‡ **Keisuke Nakamura**
長谷川雄二 ‡ **Yuji Hasegawa**
中臺一博 ‡ **Kazuhiro Nakadai**
尾形哲也 † **Tetsuya Ogata**
奥乃博 † **Hiroshi G. Okuno**

† **京都大学大学院 情報学研究科 Graduate School of Informatics, Kyoto University**
{itohara, mizumoto, angelica, ohtsuka, ogata, okuno}@kuis.kyoto-u.ac.jp
‡HRI-JP **Honda Research Institute Japan**
{keisuke, yuji.hasegawa, nakadai}@jp.honda-ri.com

Abstract

New issues will arise when plural robots participate in an ensemble with human players to attain the "unification" at three aspects of music, rhythm, melody, and harmony. We assumes that every interaction should be explicitly expressed and observable among all participants, even between robots and human. In this paper, we identify the issues for the ensemble between multiple-robot and multiple-human and we report our approach to the rhythmical unification. We focus on audio-visual integration for beat-tracking by using a Kinect with its four microphones, a stereo camera, and an infrared sensor. The resulting system provides a highly accurate beat-estimation for multi-player situations even if two robots use different beat-tracking methods.

1 はじめに

近年, 人の生活にホームロボットが密接に関わるようになっている. 人とロボットの共生において, ホームロボットが自らに付属したセンサで周囲の状況を認識し, 人と同期して動作を行うことは重要な要件である. なぜなら, 人の生活環境内で活動するロボットは, ロボット周辺の環境に対して, 適応的に人と協調作業を行う必要があるからである. この要件を満たす課題の一つに音楽合奏があげられる. 音楽合奏の課題達成は, 同期にズレを許容する表現力豊かなインタラクションの実現に重要である. また, 複数のロボット環境下でのロボット同士のコミュニケーションを, コンピュータ上の電子的なやりとりでなく人が知覚できる範囲に制限することは, 人との同期タスクを達成する上で, 必要不可欠な要素となる.

音楽は, リズム, メロディー, ハーモニーの 3 要素で定義される. 我々は, これら 3 つの要素が同期したとき合奏が成立したと定義する (3.1 節参照). 従来, 人とロボットの合奏に用いられるセンサ入力は音響信号のみであった[Otsuka, 2010; Murata, 2008]. 3 要素のうち, メロディーとハーモニーの同期は, ピッチや和音構成によって決定される. これらの要素を他のモダリティから得ることは難しく, 例えば, トロンボーンやテルミンのような手の位置によりピッチコントロールする楽器であっても, 動作が緻密であるため, 視覚から正確なピッチを得ることはできない. リズムの同期はタイミングにより決定される. タイミング検出手法として, 音楽の拍時刻とテンポを推定するビートトラッキングが盛んに研究されている. しかし従来手法の多くは, ピッチや和音同様, 入力は音響情報のみであった[Goto, 2001; Hainsworth, 2003]. その一方で人の感覚器官を考えると、発音タイミングは聴覚以外でも知覚できる。例えば, バスドラムやベースの出す低周波音の発音タイミングは触覚を用いて振動を肌で感知できる. また, 楽器演奏における発音タイミングに相関のある動作や, 演奏者同士のアイコンタクトなど, 視覚情報によるタイミング知覚は演奏者同士の同期において重要であり, 研究も盛んに行われている[Fredrickson, 1994].

我々は合奏に必要なマルチモーダルセンサとして Kinect を用いる. Kinect とは, 2010 年に Microsoft 社が発売した, RGB カメラと深度センサ, 4 チャネルのマイクアレイを搭載したデバイスである. 本来はゲーム用に発売されているが, 視覚, 聴覚, 深度のマルチモーダル情報を得られる安価なデバイスとして注目されている (2.3 節参照).

本研究では, 複数ロボット, 複数人での合奏おける, リズムの同期の部分に着目したマルチモーダル情報による合奏実現を目標とする. 複数ロボット間での同期における

課題を定義し, Kinect を用いることで, この様な複雑な実験環境下において, よりロバストなビートトラッキングを達成し, 合奏の達成を向上する.

第 2 章では音楽ロボット, ビートトラッキング, Kinect の関連研究について議論し, 本稿の立場を明らかにする. 第 3 章で複数ロボットを用いた合奏における問題を定義し, それぞれの解決に関する議論を行う. 第 4 章で上記の問題の解決の一つである視聴覚統合ビートトラッキングの Kinect を用いた視覚トラッキングについて述べ, 第 5 章で視覚トラッキングの簡単な評価, 及び複数ロボット合奏についての考察を述べる.

2 従来研究

2.1 音楽ロボットに関する研究

音楽ロボットの演奏技術の発展は近年目覚しく，発音タイミング精度や音量コントロール等による演奏表現の自由度が高いものが開発されている[Solis, 2008]．また, ロボットそのものだけではなく, 制御手法をアプローチとする研究も行われている. 水本らのロボットに依存しない汎用テルミン演奏モデル[Mizumoto, 2010a]もその一つである.

共演者ロボットの実現には，いかに協調演奏をするかという課題も存在する．Weinberg らは，ロボット 2 体と人 2 名の 4 楽器でのジャムセッションを報告している[Weinberg, 2009]．用いられたロボットは pow-wow ドラム演奏の Haile とマリンバ演奏の Shimon で，Haile はパーカッショニストとの音量主体の演奏主導権の移動を，Shimon はキーボーディストの演奏の模倣演奏と, ディスプレイを用いた疑似的なアイコンタクトとを行う．Petersen らは, フルート演奏ロボットと人のサックスとの協調演奏を報告した[Petersen, 2008]．ロボットはサックスの位置に応じて演奏パターンの変更を行う.

これらの協調演奏の対象はジャムセッションであり，タイミングに関する根本的な取り組みは行われていない．一方, ビートトラッキングという拍時刻検出手法を用いることで, 音楽に同期した足踏み[Murata, 2008]や楽譜に従った協調演奏[Mizumoto, 2010b]を行うロボット実演が報告されている. 我々もリズムに焦点をおいた合奏実現を行うため, 同様にビートトラッキングを利用する.

2.2 リズム同期に関する研究

協調演奏における要素技術の一つとして, ビートトラッキングの関連研究を示す．後藤らは, 多数のエージェントによるビートの複雑さに頑健なマルチエージェント手法を報告した[Goto, 2001]. 多数のエージェントが独立に拍時刻を推定し, 信頼度に従い分裂と消滅を繰り返す. 最終的に楽曲に一致する推定値を持つエージェントだけが残るので, 正しい拍時刻及びテンポが推定できる．村田らは, STPM(Spectro-temporal Pattern Matching) によるビートトラッキングを報告している[Murata, 2008]．STPM の利点は，定常雑音に対する頑健さ, テンポ変化に対する鋭敏さ, 実時間処理に適した動作遅延が小ささである.

パーティクルフィルタのような確率的手法を用いたビートトラッキングも報告されている．入力の特徴量として, Hainsworth らは音響信号のパワー変化[Hainsworth, 2003]を，大塚らはスペクトログラムの相互相関と楽譜情報[Otsuka, 2010]を用いている．これらの手法が音響情報のみを用いるのに対し, 我々は従来研究において, 音響情報に加え, 手のストローク動作の画像情報を用いたマルチモーダルビートトラッキングを報告した. これにより, ギター演奏という, 音がまばらで拍検出が難しい状況下でのビートトラッキングが可能になった.

画像情報を利用した他のリズム同期として, 開始及び終了タイミングの取得, 演奏主導権交代があげられる。Lim らは, Hough 変換により検出されたフルートの傾き変化に応じて, 演奏開始, テンポ変化, 演奏終了キューを検知するジェスチャー認識を報告した[Lim, 2010]. Pan らは, オプティカルフローにより顔の向きの変化を取得し, これを主導権の交代をキューの 1 つとして使用した[Pan, 2010].

2.3 Kinect, 深度情報を用いた研究

従来，深度情報を得るために，ステレオカメラや TOF(Time of Flight) カメラが用いられてきた. しかし, ステレオカメラはカメラ校正や計算コストの大きさが, TOF カメラは価格や色情報との同期の難しさが問題があった. 一方, Kinect は色情報と深度情報が紐付けされた状態で取得できる上に, 安価であるという利点がある.

Kinect の色情報と深度情報の両方を利用した研究は盛んに行われており, Saenko らは物体の高精度なラベル付け[Saenko, 2011]を, Oikonomidis らは手の関節の姿勢のトラッキング[Oikonomidis, 2011]を報告している．一方で, 音響情報を同時に利用した研究は少ない. 本稿では, 視聴覚の両方の情報を同期して用いることで, 演奏タイミングに対するより高精度な同期を行う.

3 本稿におけるロボット合奏

3.1 合奏の定義

音楽合奏は, リズム, メロディー, ハーモニーの 3 要素で構成される. リズムとは発音のタイミング, 長短, 強弱, 及びその組み合わせ, メロディーとは音の高低 (ピッチ), およびその順列 (旋律), ハーモニーはメロディーの組み合わせによる和音である.

合奏の成立を 3 つの要素が同期することと定義する. リズムの同期を各演奏者の発音タイミングの時間ズレが十分に小さい状態とする. しかし実際の合奏においては, 単に発音時間が近ければいいとは限らない. 文献[Friberg, 2002]では, ジャズ楽曲においてドラムパート, ソロパート

のスウィング, つまり基準となる時間とのズレがテンポに比例する形で現れると述べられている. 同文献によると, ソロパートにおいて, 特に表拍 (1 小節を偶数個等分したときのの奇数番目の拍) において少し遅いタイミングで演奏を行うケースが多数観測された. 同様のスウィングの研究はジャズ楽曲を中心に広く行われており, このスウィングが豊かな音楽表現につながると考えられる. メロディー及びハーモニーが同期するということはピッチ, 和音構成が同期することとみなせる. ピッチの同期とは, 2 つの音の基準音, 例えば A4 の音が十分に近いことであると定義できる. 演奏においてピッチ同期は重要であり, 多くの楽器は演奏前のピッチチューニングにより同期を行う. 一方, 管楽器のように温度の変化などでピッチが変わる場合は, 時変なピッチコントロールを行う必要がある. 単独演奏の場合, 音の相対ピッチ差が保たれていれば絶対ピッチはそれほど重要ではない. しかし, 複数による合奏演奏の場合は, パート間のピッチが近い必要がある. 一説には, 人のピッチ分解能は 5-6[cent] とされている[Loeffler, 2006]. しかし, 合奏においてピッチのズレが認知できることと, 合奏として不快と感じることとは必ずしも一致せず, この点に関する研究は不足している. 本稿では, ピッチに関する同期は使用するロボットの動作モデルに依存するため, 扱わないものとする. また, 和音の同期は主に和音の “進行”, モード (調) により決定される. セッション合奏のようなメロディーの生成が必要な場合は, 生成されたメロディーと伴奏和音の同期が特に問題となる. 本稿では演奏楽曲の和音進行とメロディーは既知であるとし, 考慮しない.

以上のことから, 合奏タスクをある程度のズレを許容したリズムの同期と定義する必要がある. 本稿では, 合奏タスクの失敗を 4 分音符間隔以上のズレが生じること, 成功を上記のような失敗が演奏中に発生しないことと定義し, 議論を進めていく.

3.2 合奏の構成

本稿の合奏の構成を, ロボット 2 体と人 2 名であるとする.

ロボットのうちは 1 台は VOCALOID による歌唱と, 手の振りによるビートタイミングに合わせたダンス[Oliveira, 2010b]を行う (以後, “ダンスロボット” と呼ぶ). もう 1 台はテルミンを演奏する (以後, “テルミンロボット” と呼ぶ). テルミンは音量とピッチの二種類のアンテナを持った非接触性の楽器である. ロボットの演奏動作には, テルミン演奏モデルに基づいた, ハードに依存する部分を分離した動作生成モジュール[Mizumoto, 2010a]を用いる. テルミンロボットのビートトラッキングの音響・画像情報の入力に Kinect を用いる. 詳しくは次節で述べるが, 音響・画像情報に加え, 深度情報を利用することで, 複数音源環境でもロバストな動作を可能とした.

人のうち一人はギターを担当する. 初期テンポ共有のために, 演奏開始時に 4 分音符間隔のギター打撃音を鳴らす. その後は楽曲に応じたストローク動作で演奏を行う.

もう一人はフルートを担当する. フルート奏者の正面に USB カメラを設置し, Ready, Start, Fermata-End の 3 つのジェスチャを検出することで, ロボットの演奏との同期を行う[Lim, 2010].

以下に合奏の構成をを示す.

合奏の構成

- ダンスロボット：歌唱 (VOCALOID)&ダンス [Oliveira, 2010b]
- テルミンロボット：テルミン演奏[Mizumoto, 2010a]
- 人：フルート (ジェスチャー認識[Lim, 2010])
- 人：ギター (ビートトラッキング)

3.3 複数ロボット合奏の問題と解決

前節で示した条件下でのマルチロボット同期合奏では, 以下のような問題が生じる.

1. 複数のビートトラッキング手法を用いた合奏遂行
2. 音源が増えたことによる検出拍候補の増加

以下でこれらの問題の解決について議論する.

3.3.1 複数のビートトラッキング手法を用いた合奏遂行

今回, 2 体のロボットそれぞれに対し, 異なるビートトラッキング手法を用いている. その理由は, ロボットのビートトラッキングに対する要求がそれぞれ異なるからである. ダンスロボットには, ダンスにおいてロボットの動作制約があるので, 大きなテンポ変動には対応できない. 一方テルミンロボットの演奏動作は, ダンスロボットに比べて比較的小さく, テンポの変動に対し機敏に対応できる. また, 担当パートがベースのような伴奏パートに当たるので, 同じく伴奏であるギターの演奏を正確に追従する必要がある. 以上より, ダンスロボットの動作タイミングは IBT[Oliveira, 2010a]による比較的人によるテンポの流動性を吸収した拍時刻を, テルミンロボットの動作タイミングは視聴覚ビートトラッキング[Itohara, 2011]によるテンポ変動に鋭敏な拍時刻を与えることとした. 以下, 前者のビートトラッキングを “ハードビートトラッキング”, 後者を “ソフトビートトラッキング” と呼ぶ.

この様な二つの異なるビートトラッキング手法間での同期を解決する方法は二つある. 一つはロボット間で電子的な通信を行うことである. しかし, これは人にはその同期の様子が伝わらず, 人との同期の要件を満たすことはできない.

もう一つの方法は, 人同士の合奏同期と同様に “リズムリーダー” を決めることである. オーケストラで言えば指揮者が, ロックバンドで言えばドラム奏者がそれにあた

る．リーダーとの同期は視覚や聴覚と言った様々なモダリティで行われている．本稿における合奏では，ギター奏者がリーダーにあたる．しかし，人の，特にアマチュア奏者の演奏の場合，テンポの流動性は回避しきれない．その一方でダンスロボットは，ハードビートトラッキングを用いており，また，ダンスや歌声という視聴覚からリズムのとりやすい動作をしている．よって，ダンスロボットを相補的なリズムリーダーとすることで，アマチュア奏者でもロボットとのテンポの安定した同期合奏が実現できる．

3.3.2　音源の増加による検出拍候補の増加

テルミンロボットはギター奏者に追従するビートトラッキング手法を用いている．しかし，テルミンロボットの入力デバイスである Kinect には，ギター以外の多数の音が混合された状態で入力される．各演奏者の発音タイミングにはズレが生じているので，ソフトビートトラッキングがギター以外のズレに引きずられ，その誤差が蓄積することで，合奏タスクが失敗する可能性がある．これに対する一般的な解決法は，(1) 対象楽器に対するトラッキングのモダリティを増やして精度を高めること，(2) 方向による音源分離をしてギターの音だけを強調すること，である．(1) は，リズムリーダーの音だけでなく，動きなど別の要素に注目することで同期を図ることに一致する．(2) はカクテルパーティー効果のような，多数ある音の中から目的音だけに着目することと同義である．本稿では (1) を採用する．具体的には，Kinect の深度情報を使い，ギターの視聴覚統合ビートトラッキングの追従性能を向上させることで上記の問題を解決する．次章にてその詳細を示す．

4　Kinect を用いた視聴覚統合ビートトラッキング

4.1　視聴覚統合ビートトラッキング概要

本稿では，ロボットがギター演奏との同期を行うために，ギターの演奏音と手の動作との相関性を用いた視聴覚統合ビートトラッキング[Itohara, 2011]を用い，ロボットの演奏タイミング検出を行う．出力は入力演奏のテンポと 1 小節を 4 等分した拍の位置であり，ロボットはこれで示されるタイミングに基づいて演奏する．ロボットの演奏動作生成に時間がかかるので，それにしたがって拍推定は少し時間を遡って行われる．本実験では 500[msec] とした．

従来手法において，視覚情報処理，つまり手のトラッキングの部分で，手とギターの色が似ているために起こる手の誤検出の問題があった．本章では，Kinect の深度情報を用いたギター平面の検出，及び画像マスキングによる，色の類似に頑健な手のトラッキングを報告する．これにより，ギタービートトラッキングの性能向上が期待される．音響情報・視覚情報処理，及びパーティクルフィルタの実装に関する詳細は文献[Itohara, 2011]に譲り，以下ではギターのマスキングについて述べる．

4.2　深度情報によるギターマスキング

Kinect による入力は，サイズ 640×480[pixel] の RGB と深度画像である．また，深度画像として各ピクセル座標における x, y, z 方向の値 (単位:[m]) が得られる．x, y, z 正の方向はそれぞれ水平 Kinect からむかって左，鉛直下向き，カメラ方向である．

以下にギターのマスキングの過程を示す．

1. 背景閾値以上の奥行き (z) を持つ座標を，RGB 画像，深度画像においてマスクをかける
2. 深度画像を縮小．以下，非マスク部を “特徴点” とする．
3. 特徴点からギターの平面パラメータを導出
4. 深度画像の各座標と 3 の平面の距離を計算し，閾値以下なら対応する RGB 画像上の点にマスクをかける

本稿では，背景閾値を 3[m]，画像の量子化は 16×12，4. の平面との距離閾値は 5[cm] とした．以下で 3. における，3 次元空間における Hough 変換を用いた，ギターの表板を表す平面パラメータの推定について述べる．

Hough 変換では，画像中の各特徴点を通るすべての平面のパラメータを算出，パラメータ空間に対して投票を行い，最大票を獲得したパラメータを推定平面のパラメータであるとする．平面パラメータは，球座標における原点を始点とする平面の法線ベクトル (ρ, θ, ϕ) である．ρ は原点と平面の距離を表す．それぞれのパラメータの定義域，各パラメータの空間の分割幅は以下のように定めた．

$$0.7 \leq \rho \leq 1.4, \qquad \rho_{bin} = 0.05[\mathrm{m}]; \tag{1}$$

$$0 \leq \theta < \pi/4, \qquad \theta_{bin} = \pi/16[\mathrm{rad}]; \tag{2}$$

$$0 \leq \phi < 2\pi, \qquad \phi_{bin} = \pi/6[\mathrm{rad}]; \tag{3}$$

5　実験検証

本章では，4.2 節での深度情報を用いた手のトラッキングの簡単な評価を行う．また，それらを実際に適応した複数ロボット，複数人数合奏実験の検証を行い，解決された課題，今後に残された課題についての議論を行う．

5.1　ギターマスキングによる手のトラッキング

本節では，深度情報を用いた手のトラッキングを RGB 入力のみのものと比較し，性能比較を行う．二つの手のトラッキング手法は，入力がマスクされた画像か否か以外は同じで，オプティカルフローにより変位ベクトルをとった後，その平均を中心とした矩形と色相カーネルを用いた平均値シフト法により手の位置を取得する．詳しくは，文献[Itohara, 2011]を参照されたい．

図 1 に手のトラッキングの結果の一部を示す．既存の RGB 画像のみの入力では，手を指し示すカーソルがギター

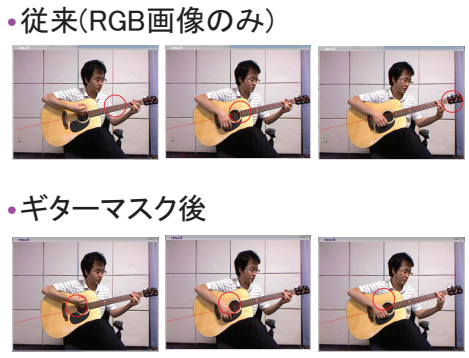

Figure 1: 手のトラッキング結果の比較．赤い丸が手の位置に対応する．上段と下段は同じフレームに対応している．

Figure 2: 合奏デモの様子．右側が Hearbo(1 号機) で，左側が Hearbo(2 号機)．

に吸い寄せられ手を避けるかのような挙動が見られることがあった．これは色相カーネルの選択において，色相の近いギターの色に引き寄せられたことが原因だと考えられる．また，ギターのヘッド (ギター左手側の先) 部分も手同様演奏中に動くことがある．これによりオプティカルフローベクトルがそちらに現れ，ヘッド部分にトラッキングカーソルが動く場合が見られた．一方，マスク後は上記のような誤検出は一切見られなかった．ただし，計算コストがフレームレートに比べ大きすぎるため，手のトラッキング結果が表示されるころには実際の手の位置が異なってしまっていることが確認できた．

5.2 複数ロボットによる同期合奏

3.2 節で示した構成で合奏デモを行なった．使用したロボットは，HRI-JP のヒューマノイドロボット，Hearbo(1 号機，2 号機) で，肩，肘，手首，指 (片手)，首にそれぞれ 2,1,2,4,3 の自由度を持つ．Hearbo(1 号機) にダンスロボットを，Hearbo(2 号機) にテルミンロボットを割り当てた．楽曲はイングランド民謡のグリーンスリーブスを用いた．図 2 に実際の合奏の写真を示す．前節の手のトラッキング精度向上の結果，テルミンロボットのギターへの追従性は大きくあがったと言える．また，ダンスロボットのハードビートトラッキングにより，人のテンポ流動性が抑えられ，ズレの誤差の蓄積が減り，合奏タスクの成功率が向上した．今後はこれらのタスク成功率を定量的に評価する必要がある．

今回の合奏実験の成功率は 30%程度に留まっている．この一番の原因は，2 つのビートトラッキング手法間の同期が失敗することにある．これを解決するためには，ロボット間の明示的な同期が必要である．人同士の場合は，リズムリーダーを最も信頼をした拍推定を行なって同期を行なうが，同時に全体の同期も考慮する．この行動は，例えば誰かがフレーズを誤った場合にそちらを注視するといった行動に現れる．現在のビートトラッキング手法では，どちらも一つの音源のみに頼ったビートトラッキングを行なっており，ロボット同士はお互いの演奏や動作を一切考慮していない．この状況が解決されることで，合奏の成功率は大きく上がると考えられる．

6 おわりに

本稿では，複数ロボットと複数人で構成されるリズム同期に焦点をあてた合奏について議論し，ロボット 2 体と人 2 名による合奏システムについて報告した．その際，リズム同期で必要なビートトラッキングにおいて，ハードビートトラッキングを用いることで，人の流動的な演奏と合わせて相補的なテンポの維持が可能になった．また，ソフトビートトラッキングにおいて，Kinect の深度情報を合わせたマルチモーダルビートトラッキングを用いることで，伴奏演奏により合った同期演奏が達成できた．

今後の課題として，ロボット合奏タスク達成率の向上があげられる．そのためには，ビートトラッキング自体の精度向上，複数のビートトラッキング手法間の協調のようなロボット間の同期の実現が必要となる．本稿では，リズムのみを考慮した合奏を行なったが，ピッチなどの他の要素が同期ことに注目した合奏考察は，豊かなインタラクションという点で重要である．例えばピッチにおいては，声楽における長音のビブラート，ギターなどの弦楽器で弦を押し上げることによるピッチ変化などに現れる．また，テルミンのような無音階楽器の演奏における他の楽器とのピッチ同期は特に重要である．これに対し，実際の演奏音を聞いて，適応的に動作を変化させるロボット演奏の研究も行なわれている[水本他, 2011]．また，従来の音楽合奏タスクの評価は，本稿同様，リズムのみに着目したものが多かった．上記のようなピッチ変動，またはセッションなどのメロディー生成における和音同期や，合奏としての楽しさといった主観的な要素などを盛り込んだ，新しい合奏の評価尺度の考察が必要である．

謝辞 本研究の一部は科研費 (S)，新学術領域，JST-ANR BINAAHR，GCOE の支援を受けた．また，Hearbo の使用許可をいただいた HRI-JP に感謝します．

参考文献

[Fredrickson, 1994] W.E. Fredrickson. Band musicians' performance and eye contact as influenced by loss of a visual and/or aural stimulus. *Journal of Research in Music Education*, 42(4):306, 1994.

[Friberg, 2002] A. Friberg and A. Sundström. Swing ratios and ensemble timing in jazz performance: Evidence for a common rhythmic pattern. *Music Perception*, 19(3):333–349, 2002.

[Goto, 2001] M. Goto. An audio-based real-time beat tracking system for music with or without drum-sounds. *J. of New Music Research*, pages 159–171, 2001.

[Hainsworth, 2003] S. Hainsworth and M. Macleod. Beat tracking with particle filtering algorithms. In *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,*, pages 91–94. IEEE, 2003.

[Itohara, 2011] T. Itohara et al. Particle-filter based audio-visual beat-tracking for music robot ensemble with human guitarist. In *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*. IEEE, 2011.

[Lim, 2010] A. Lim et al. Robot musical accompaniment: integrating audio and visual cues for real-time synchronization with a human flutist. In *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, pages 1964–1969, 2010.

[Loeffler, 2006] B.D. Loeffler. *Instrument Timbres and Pitch Estimation in Polyphonic Music*. PhD thesis, Citeseer, 2006.

[Mizumoto, 2010a] T. Mizumoto et al. Human-robot ensemble between robot thereminist and human percussionist using coupled oscillator model. In *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, pages 1957–1963. IEEE, 2010.

[Mizumoto, 2010b] T. Mizumoto et al. Integration of flutist gesture recognition and beat tracking for human-robot ensemble. In *Proc. of IEEE/RSJ-2010 Workshop on Robots and Musical Expression*, pages 159–171, 2010.

[Murata, 2008] K. Murata et al. A beat-tracking robot for human-robot interaction and its evaluation. In *Proc. of 8th IEEE-RAS Int'l Conf. on Humanoids*, pages 79–84. IEEE, 2008.

[Oikonomidis, 2011] I. Oikonomidis et al. Efficient model-based 3d tracking of hand articulations using kinect. *Procs. of BMVC, Dundee, UK (August 29–September 10 2011)[547]*, 2011.

[Oliveira, 2010a] J.L. Oliveira et al. Ibt: A real-time tempo and beat tracking system. In *Proc. of Int'l Society for Musical Information Retrieval Conference*. IEEE, 2010.

[Oliveira, 2010b] J.L. Oliveira et al. Synthesis of dancing motions based on a compact topological representation of dance styles. In *Proc. of IEEE/RSJ-2010 Workshop on Robots and Musical Expression*. IEEE/RSJ, 2010.

[Otsuka, 2010] T. Otsuka et al. Design and Implementation of Two-level Synchronization for Interactive Music Robot. In *Proc. of Association for the Advancement of Artificial Intelligence*, pages 1238–1244, 2010.

[Pan, 2010] Y. Pan et al. A robot musician interacting with a human partner through initiative exchange. In *Proc. of Int'l Conf. on New Interfaces of Musical Expression*, pages 166–169, 2010.

[Petersen, 2008] K. Petersen et al. Development of a real-time instrument tracking system for enabling the musical interaction with the waseda flutist robot. In *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, pages 313–318, 2008.

[Saenko, 2011] K. Saenko et al. Practical 3-d object detection using category and instance-level appearance models. In *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*. IEEE, 2011.

[Solis, 2008] J. Solis et al. Understanding the mechanisms of the human motor control by imitating flute playing with the Waseda Flutist Robot WF-4RIV. *Mechanism and Machine Theory*, 44(3):527–540, 2008.

[Weinberg, 2009] G. Weinberg et al. The Creation of a Multi-Human, Multi-Robot Interactive Jam Session. In *Proc. of Int'l Conf. on New Interfaces of Musical Expression*, pages 70–73, 2009.

[水本他, 2011] 水本 武志 他. テルミン演奏ロボットのための unscented kalman filter による適応的音高制御. In **日本ロボット学会第 *29* 回学術講演会**, 2011.