Vol.2012-MUS-94 No.6
Vol.2012-SLP-90 No.6
2012/2/3

VocaWatcher: 人間の歌唱時の表情を真似るヒューマノイドロボットの顔動作生成システム

中野倫靖[†1] 後藤真孝[†1] 梶田秀司[†1]
松坂要佐[†1] 中岡慎一郎[†1] 横井一仁[†1]

本稿では、人間の歌唱における顔表情を真似てロボットの顔動作を生成する VocaWatcher について述べる。歌声は、我々が以前開発した VocaListener で合成する。従来、歌唱ロボットに関する研究はあったが、手作業による動作制御が主で、その自然さに限界があった。それに対して本研究では、単一のビデオカメラで収録した人間の歌唱動画を画像解析し、口、目、首の動作を真似て制御することで、自然な歌唱動作を生成した。ここで口の制御には、VocaListener から得られる歌詞のタイミング情報を用いて、歌声に同期した動作を生成できる。さらに、既存の VocaListener を、ブレス音を真似るように拡張して合成することで、より自然なロボット歌唱を実現した。

VocaWatcher: A Facial-Motion Generation System for Humanoid Robot by Imitating Facial Expressions of Human Singer

Tomoyasu Nakano,[†1] Masataka Goto,[†1]
Shuuji Kajita,[†1] Yosuke Matsusaka,[†1]
Shin'ichiro Nakaoka[†1] and Kazuhito Yokoi [†1]

In this paper, we describe *VocaWatcher* that is a facial-motion generator for singing robot by imitating human movements. It can also synthesize singing voices by using our previous VocaListener to imitate human singing. Although singing humanoid robots have been developed with synthesized singing voices, such robots do not appear to be natural because of the limitations of manual control. To generate natural singing expressions, VocaWatcher imitates a human singer by analyzing a video clip of a human singing, recorded by a single video camera. VocaWatcher can control mouth, eye, and neck motions by imitating the corresponding human movements. To control the mouth motion, VocaWatcher uses lyrics with precise timing information provided by VocaListener. Moreover, we extended VocaListener to imitate breathing sounds that make the robot singing more realistic.

1. はじめに

本研究では、ヒューマノイドロボットの自然な歌唱動作の実現を目指し、その第一段階として、人間の歌唱を収録した動画像を入力として、ヒューマノイドロボットがその歌い方（歌い回しと顔表情）を真似て歌うための、歌声合成技術及び顔動作生成技術について述べる。経済産業省「技術戦略マップ 2010」[⋆1]における「アイドルロボット」構想（ソフト：コンテンツ分野）に見られるように、ヒューマノイドロボットは多くの人々の関心を惹き付けやすく、ロボット技術のエンターテインメント分野への展開は、その関心の高さに裏付けられた有望な応用事例である。そのようなエンターテインメント応用に向けた、ヒューマノイドロボットとの親和性が高い技術の一つに歌声合成がある。

2007 年以降、日本では市販の歌声合成ソフトウェアが注目を集め、それを用いて楽曲制作を行う一般ユーザや音楽家（プロ）が急増してきた。そうして創られた作品の多くは、動画コミュニケーションサービス「ニコニコ動画[⋆2]」に投稿されたことで、多数のリスナーによって鑑賞されるだけでなく、そのコンテンツの一部、もしくは全部が新しいコンテンツの中で再利用されるといった、Web を介した大規模な協調的創造活動につながってきた[1),2)]。また、そうした楽曲が収録された音楽 CD が市販され、商業音楽ヒットチャート上位にランクインするなど、歌声合成による楽曲を楽しむリスナーも増えている[3)]。さらに最近では、CG 映像によるライブが日本国内や海外で開催される[⋆3]など、音楽の鑑賞における新たな形態も登場し始めている。過去にシンセサイザが新たな楽器として普及してきたように、新たな歌声としての歌声合成が社会的に普及しつつあることから、歌声合成によって歌うヒューマノイドロボットは、高い応用可能性があると考えられる。

これまで、我々は人間に近い外観[⋆4]と動作性能を持ち、表情制御可能なヒューマノイドロボット HRP-4C[4)] を開発してきた（**図 1**）。さらに、ヤマハ株式会社と共同で、歌声合成ソフトウェア Vocaloid2[3),5)] を用いて、HRP-4C に歌唱させる展示も行うなど[6)]、ヒューマノイドロボット技術のエンターテインメント応用への可能性を検討してきた。しかし、顔動作

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

⋆1 http://www.meti.go.jp/policy/economy/gijutsu_kakushin/kenkyu_kaihatu/str2010.html
⋆2 http://www.nicovideo.jp/
⋆3 現時点では、アメリカとシンガポールにて開催。
⋆4 身長 160cm、体重 46kg（バッテリー含む）、44 自由度で、関節位置や寸法は日本人青年女性の平均値を参考に、人間に近い外観を実現した。

Vol.2012-MUS-94 No.6
Vol.2012-SLP-90 No.6
2012/2/3

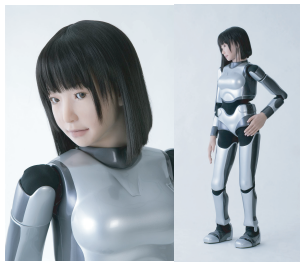

図 1 ヒューマノイドロボット HRP-4C の外観

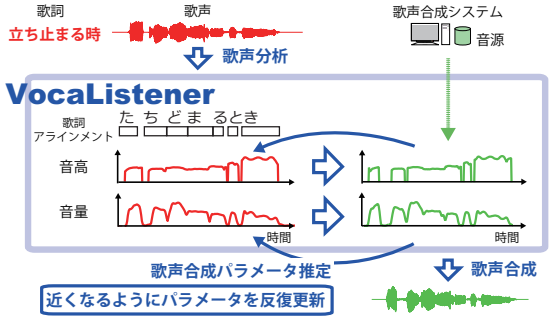

図 2 VocaListener の処理概要。人間の歌声と歌詞を入力として、その歌い方に近くなるように歌声合成パラメータを反復推定して歌声合成する。

の生成と歌声の合成は、事前に用意したテンプレートの状態遷移モデルやルールベースの制御、手作業によって行っていたため、その表現力には限界があった。

そこで本研究では、歌唱の表現力向上のために、人間の歌い方を真似して歌声合成する既存のシステム VocaListener[7),8)]（**図 2**）を導入して歌声を合成した。さらに、VocaListener と同様の枠組みに基づく顔動作生成システム **VocaWatcher**[9),10)] を新たに実現し、単一の家庭用ビデオカメラで撮影された人間の歌い手の映像を用いて、その顔表情を真似るようにヒューマノイドロボットの顔動作を生成した。ここで口の制御には、VocaListener から得られる歌詞のタイミング情報を用いて、歌声に同期した動作を生成できる。さらに、人間の顔表情を真似る過程で、息継ぎで息を吸う動作と共にブレス（吸気）音の合成が必要になったので、既存の VocaListener をブレス音を真似るように拡張して合成する。

2. 関 連 研 究

ヒューマノイドロボット研究の音楽への展開は、WABOT-2 の電子オルガン演奏[11)] から始まり、フルート演奏[12)]、テルミン演奏[13)] 等が存在する。また、歌を歌わせる試みとしては、声道モデルの機械系による実現とその計算機制御による歌声合成[14)]、アカペラ歌唱やダンス可能なロボット[15)]、リアルタイムビートトラッキング技術に基づいて拍に合わせて歌って踊るロボット[16)]、簡略化された楽譜映像を認識して歌う顔ロボット[17)]、等が研究されてきた。しかし、表情制御に関してはハードウェアの制約から、十分な検討がなされていなかったり、自然な顔動作の生成や歌声の合成ができなかった。

一方、音楽や歌唱以外では、人間の顔動作に基づいたヒューマノイドロボットの制御とし

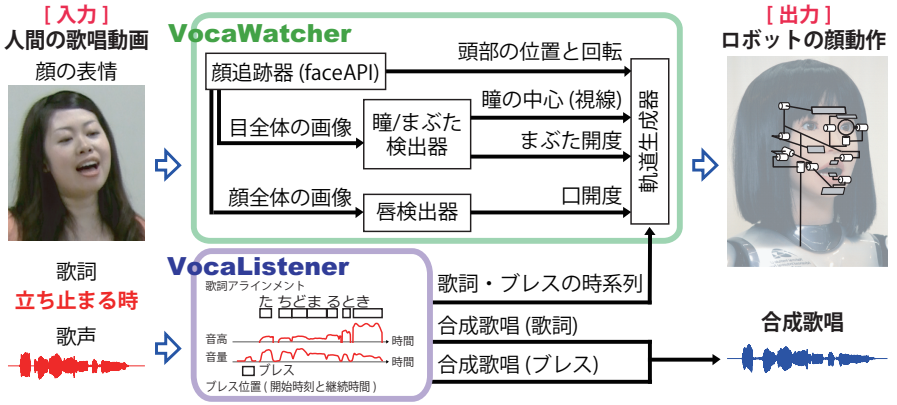

図 3 VocaListener 及び VocaWatcher による、人間の歌唱を真似るヒューマノイドロボット動作制御の処理概要

て、モーションキャプチャ結果[18)] や、人のビデオ映像の顔追跡結果[19)] を入力として用いる研究がある。しかしこれらは、顔へのマーカー付与が必要であったり[18)]、多くの学習とチューニングを要する[19)] など、我々の目的に合致した手法ではなかった。

また、歌唱における歌声と顔の表情（特に、歌詞の音素と口の形状）の間には、密接な関係があるが、歌声情報処理[20),21)] を顔動作制御に組み合わせた例はなかった。

3. 人間の歌い方を真似るヒューマノイドロボットの処理概要

本研究では、「人間の歌唱の模倣」によってヒューマノイドロボットの歌唱動作生成を実現する（**図 3**）。その機能は、人間の歌い方を真似て歌声合成する VocaListener と、人間の顔表情を真似て顔動作生成する VocaWatcher から構成される。ここで、歌唱者が自由に表現できるよう、歌唱の収録にはマーカーや視線計測器などの特別な機器は用いず、単一のビデオカメラによる動画のみを用いる。

VocaListener は、既存の歌声合成ソフトウェア (例えば Vocaloid[5)]) の歌声合成パラメータを、ユーザ歌唱からその音高と音量を真似て推定する技術である (図 2)[7),8)]。パラメータの反復推定により、推定精度が従来研究[22)] に比べて向上し、歌声合成システムやその音源（歌手の声）を切り替えても再調整せずに自動的に合成できる[⋆1]。独自の歌声専用音響モデ

⋆1 合成結果の具体例は、ホームページ http://staff.aist.go.jp/t.nakano/VocaListener/ や動画コミュニケーションサービス『ニコニコ動画』http://www.nicovideo.jp/mylist/7012071 上で視聴できる。

Vol.2012-MUS-94 No.6
Vol.2012-SLP-90 No.6
2012/2/3

図 4 歌唱収録風景

ルによって、歌詞のテキストを歌詞を音符毎に割り当てる作業は、ほぼ自動で行える⋆1。ここで本研究では、ブレス音を自動検出して、それを真似るように合成する拡張を行った。

一方、VocaWatcher には、人間の歌唱映像と VocaListener によって分析された歌詞の音節（モーラ）の時刻情報⋆2を入力として与え、「頭部の位置と回転」、「まぶた開度」、「口開度」、「視線の方向」、「唇形状」を制御する。ここで、口開度と唇形状については、VocaListener から同時に得られる発音のタイミング情報に基づいて、歌声に同期した動作を生成する。

人間の歌唱収録の様子を**図 4** に示す。左端のカメラで撮影された上半身のビデオ画像とマイクにより収録された歌声を用いた。ここで、映像は 1920 × 1080（29.97FPS）で収録したが、VocaWatcher では、その解像度を全て 960 × 540 にリサンプリングして使用した。ここで対象とする楽曲には、RWC 研究用音楽データベース（ポピュラー音楽）[23] の「PROLOGUE」(RWC-MDB-P-2001 No.7) を使用して、日本人女性 1 名による歌唱を収録した。また歌声合成システムとしては「Vocaloid2 初音ミク⋆3」を用いた。

4. 人間の歌唱に基づく顔動作生成システム VocaWatcher

本章では、新規開発した VocaWatcher について、技術上の課題と解決方法を説明する。VocaWatcher は、撮影された動画からロボットの顔動作制御のための値を推定する「人間の歌い手の顔表情分析」（**4.1** 節）と、その分析結果をロボットの顔動作制御パラメータとして実現する「ヒューマノイドロボットの顔動作生成」（**4.2** 節）で構成される。

4.1 人間の歌い手の顔表情分析

従来、瞳検出に関する研究は、視線検出に基づく車いす制御[24] や、まばたき検出に基づくコマンド制御[25] などの応用がなされてきた。しかし、歌唱中の感情表現には「半目で歌う」、「ゆっくり瞳を開く」などの連続的な変化をするため、従来技術のような離散的な開閉判別のみでは対処しきれず、まぶたの連続的な変化に対応できる手法が必要となる。

また視線検出では、できるだけ高解像度な目の画像が得られることが望ましい。しかし、歌唱中の人間は歌唱動作として常に頭を動かす傾向にあるため、動画中の全フレームにおいて顔を捉える必要がある。したがって、離れた位置から撮影した映像しか用いることができず、そのような遠い（低解像度な）目の画像から瞳（視線）を検出しなければならない。

以上の問題を解決する手法について、本節で以降、説明する。

4.1.1 頭部の位置と回転の検出（顔追跡）

顔表情分析の最初のステップとして、三次元空間における頭部の位置と回転（姿勢）を推定する。本稿では、Seeing Machine 社の顔画像トラッキングソフトウェア faceAPI[26] を用いて、各映像フレームにおける頭部の姿勢（ロール角、ピッチ角、ヨー角）と顔の特徴点（Face landmarks）の座標を得る。**図 5** に歌唱動画から推定された 1 曲（298.2 秒）中の頭部の姿勢、**図 6** に検出された特徴点の例を示す。

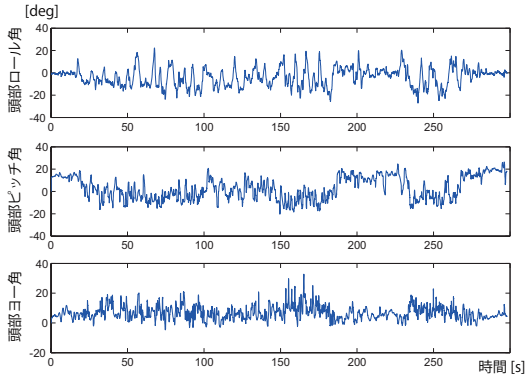

図 5 画像から推定された歌い手の頭部の動き

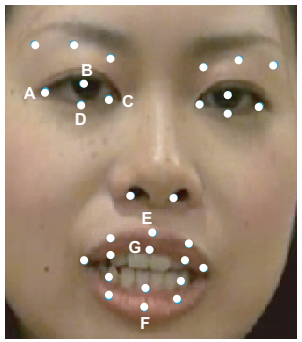

図 6 検出された顔の特徴点

4.1.2 瞳検出、まぶた検出

前述したように、歌唱中の人間の顔表情には、感情表現として半目を開くなどの連続的な動きが存在するため、通常の方法では安定した瞳の検出が困難であった。例えば、図 6 において、点 A,B,C,D で囲まれた領域が右目に対応するが、現状で用いている faceAPI (FaceTrackingAPI 3.2) では、まばたきを検出できず、目を閉じた場合でも点 B,D 間の距

⋆1 音符の割り当てでは、その推定時刻に誤りが発生する可能性があるが、誤った箇所を指摘して「ダメ出し」するだけで、新しい候補を再提示する機能もある。
⋆2 ここで各モーラの開始時刻は、母音（/t a/ であれば/a/部分）の開始時刻が出力される。
⋆3 http://www.crypton.co.jp/mp/pages/prod/vocaloid/cv01.jsp

Vol.2012-MUS-94 No.6
Vol.2012-SLP-90 No.6
2012/2/3

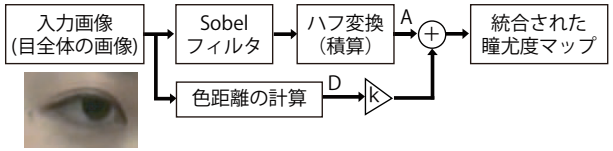

図 7 領域内の色による重みを加えたハフ変換に基づく瞳検出の概要

離が変化しないという問題があった。そこで、瞳（視線）とまぶた（まばたき）の検出には、faceAPI によって検出された目領域の画像に対して、それぞれ以下の処理を適用する。

領域内の色による重みを加えたハフ変換（瞳検出）：

図 7 に瞳検出の概要を示す。Sobel フィルタによるエッジ画像にハフ変換を行い、領域内の色の重みを加えることで検出結果を頑健にする。具体的には、二次元画像の座標を x, y とした時に、円形ハフ変換による投票結果を $A(x,y)$、色距離から算出した尤度マップを $D(x,y)$、重み付け定数を k として、瞳の尤度マップ $L(x,y)$ を以下の式から算出する。

$$L(x,y) = A(x,y) + k \cdot D(x,y) \tag{1}$$

ここで、$A(x,y)$ が形で $D(x,y)$ が色を手がかりとした瞳の存在確率に相当し、手がかりを増やして頑健性の向上を図っており、入力画像を $I(x,y)$、Sobel 演算子によって得られるエッジ情報を $|\nabla I(x,y)|$、瞳領域の輝度を I^r、p_r と θ を円形ハフ変換における円の半径（原点からの距離）と角度、としてそれぞれ次のように算出される。

$$A(x,y) \leftarrow A(x,y) + |\nabla I(h_x, h_y)| \tag{2}$$
$$(h_x = p_r \cos\theta + x, h_y = p_r \sin\theta + y)$$
$$|\nabla I(x,y)| = (dI/dx + dI/dy)^{1/2} \tag{3}$$
$$D(x,y) = (1 - I(x,y) - I^r)^2, \tag{4}$$

本稿では、歌い手が日本人であるため、瞳は黒と仮定して色距離 $D(x,y)$ はモノクロ画像から算出し、式 (2) では座標 h_x, h_y のピクセル（画素値）が瞳の円周上の境界（エッジ）だった場合に、より大きな値でハフ変換用に積算されるここで、変数 θ を一周分変化させながら、瞳の中心 x, y に対して積算値を $A(x,y)$ として記録している。円の半径 p_r については、目領域の高さから想定される半径の値の範囲について、各ハフ変換と投票結果を計算（半径の大きさで正規化）し、最も投票が多かった候補を最終的な瞳の半径とした。

このようにして得られた瞳の尤度マップ $L(x,y)$ から、瞳の位置 p_x, p_y（それぞれ x 軸と y 軸における値）を次のように決定した。

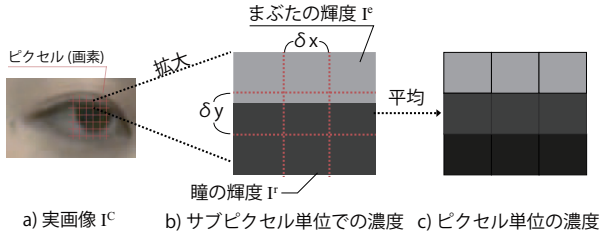

図 8 サブピクセル情報を用いた目領域の分解能向上（まぶた検出）におけるピクセルと実画像の関係

$$(p_x, p_y) = \underset{x,y}{\operatorname{argmax}}\, L(x,y) \tag{5}$$

サブピクセル情報を用いた目領域の分解能向上（まぶた検出）：

前述した頭部全体を撮影する必要性から、目領域の解像度は少なく 3〜6 [pixel] であった。通常のピクセルベースの検出では、まぶた開度に 3〜6 の離散値しか得られず、歌唱表現を適切に反映できないため、サブピクセル情報を用いて分解能をあげて処理を行う（**図 8**）。

連続領域における実物体が発する輝度を $I^C(x,y)$、ピクセルの幅と高さを δx と δy とすると、標本化して観測される各ピクセルの輝度 $I(\bar{x}, \bar{y})$ は以下の式の関係になると仮定できる。

$$I(\bar{x}, \bar{y}) = \frac{\int_{\bar{y}-\frac{\delta y}{2}}^{\bar{y}+\frac{\delta y}{2}} \int_{\bar{x}-\frac{\delta x}{2}}^{\bar{x}+\frac{\delta x}{2}} I^C(x,y)dxdy}{\delta x \delta y} \tag{6}$$

ここで、前節の式 (5) で得られた瞳の x 軸方向の中心位置 p_x を利用し、その中心位置を通る垂直線上（y 軸に平行な線上）でのまぶたの境界位置を b（y 軸方向の位置）とする。この b が含まれるピクセル、つまり、まぶたと瞳の境界領域にあるピクセルに着目して、上記の輝度の式を用いたい。そのために、b より上のまぶたの輝度が I^e、b より下の瞳の輝度が I^r で一定であると仮定し、そのピクセルの y 軸方向の位置を B_y とすると、その輝度 $I(p_x, B_y)$ は面積に応じた重み付け和として次のように近似できる。

$$I(p_x, B_y) = \frac{\int_b^{B_y+\frac{\delta y}{2}} \int_{p_x-\frac{\delta x}{2}}^{p_x+\frac{\delta x}{2}} I^e dxdy + \int_{B_y-\frac{\delta y}{2}}^{b} \int_{p_x-\frac{\delta x}{2}}^{p_x+\frac{\delta x}{2}} I^r dxdy}{\delta x \delta y} \tag{7}$$
$$= \frac{(B_y + \frac{\delta y}{2} - b)I^e + (b - (B_y - \frac{\delta y}{2}))I^r}{\delta y} \tag{8}$$

これを変形して、b は次のように求まる。

Vol.2012-MUS-94 No.6
Vol.2012-SLP-90 No.6
2012/2/3

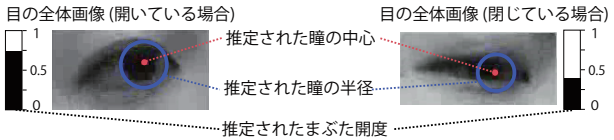

図 9 瞳の中心と半径、まぶた開度の推定結果の例

$$b = \frac{\delta y I(p_x, B_y) - (B_y + \delta y/2) I^e + (B_y - \delta y/2) I^r}{I^r - I^e} \tag{9}$$

ただし、現在の実装では、前節で得られた瞳の半径 p_r と、式 (5) で得られた瞳の y 軸方向の中心位置 p_y を利用し、上記のサブピクセルの考え方を用いて、瞳全体があたかも一つのピクセル（中心位置が p_y、縦方向の長さが $2p_r$ のピクセル）であるかのように単純化することで、まぶたの開度 a を以下の式で求めた。

$$a = \begin{cases} 0 & (e < e_{min}) \\ \frac{e - e_{min}}{e_{max} - e_{min}} & (e_{min} \le e < e_{max}) \\ 1 & (e \ge e_{max}) \end{cases} \tag{10}$$

$$e := \sum_{\bar{y}=p_y-p_r}^{p_y+p_r} I(p_x, \bar{y}), \quad e_{min} := 2p_r I^e, \quad e_{max} := 2p_r I^r$$

ここで、e は瞳全体を大きなピクセルとみなした輝度に相当し、標本化して観測された瞳の画素値を $\sum$ によって瞳の直径分だけ加算して求めた。瞳の輝度 I^r は定数とし、まぶたの輝度 I^e は瞳の範囲から外れていると考えられる目領域の境界周辺のピクセルの輝度値の平均をとることで算出した。

以上の処理によって得られた、瞳の位置と半径、まぶた開度の例を**図 9** に示す。

4.1.3 口開度の検出

歌唱時の高速な唇の動きのために faceAPI はしばしば唇のトラッキングに失敗し、正確な口開度（上唇と下唇間の距離）を検出できなかった。そこで、まず faceAPI で得られた特徴点で定められる顔の中心線 (図 6 において、線分 E-F に平行で点 G を通る直線) に沿った一次元のイメージを元画像より抽出し、時間軸に沿って並べた二次元イメージを作成した (**図 10**(a))。ここで、上下の唇はほぼ等しい色をもった帯として表れている。その時間変位を得るため、RGB の色距離を用いたパーティクルフィルタによって、上唇と下唇の中心線を推定した。このようにして得られる唇の距離を $[0, 1]$ で正規化して、口開度 c とした。

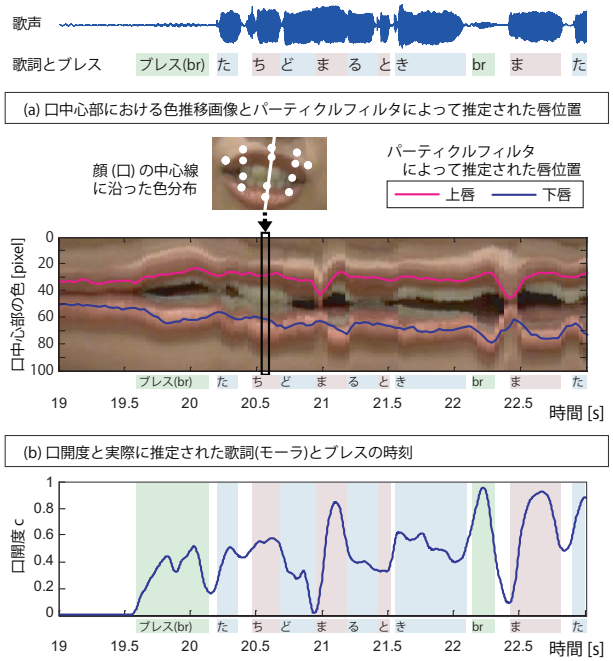

図 10 (a) パーティクルフィルタによって推定された唇の動き、(b) 口開度と VocaListener によって推定された歌詞とブレスの時刻の比較

図 10(b) に、歌い出しにおける口開度と、実際に歌われた歌声、そして VocaListener で得られた歌詞とブレスの時刻情報を比較して示す。

4.2 ヒューマノイドロボット HRP-4C の顔動作生成

前節までで、人間の顔表情データとして瞳位置、まぶた開度、口開度、歌声情報として歌詞とブレスの時刻情報が推定できたため、それに基づいてロボットの関節軌道（制御パラメータ）を推定する。**図 11** に HRP-4C の頭部の関節軸構成を示す[27]。ここで、それぞれの関節角をサーボモータにより 5ms の時間分解能で制御することで、顔動作を生成する。

4.2.1 首動作の生成

ロボットの首関節（NECK_R, NECK_P, NECK_Y）の制御は、顔動作分析において、29.97FPS で得られた頭部ロール角、ピッチ角、ヨー角 [deg] の時系列データ（図 5）を用いる。モータ制御に合わせ、5ms の時間分解能に線形補間してリサンプリングするが、そ

Vol.2012-MUS-94 No.6
Vol.2012-SLP-90 No.6
2012/2/3

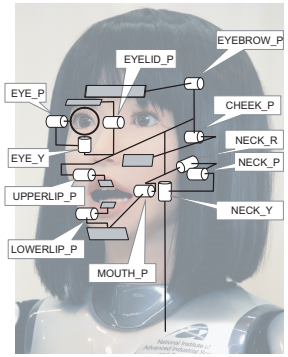

図 11 HRP-4C の顔と首の関節軸構成[27]。円柱がモータ、平行四辺形が皮膚を変形させるための機構の動作端、右目の円は眼球を示す。「それぞれの関節名の末尾で、制御可能な回転軸方向を示しており、「_R(ロール軸)」「_P(ピッチ軸)」「_Y（ヨー軸)」である。

の際には、モータ性能を考慮して、動作速度と動作範囲の抑制のために、カットオフ周波数 4 Hz のローパスフィルタ（2 次バタワースフィルタ）と、スケーリング（現在は、ゲインとして 0.6 を用いた）を施して生成した。

4.2.2 視線・まばたき動作の生成

視線やまばたきなどの関節軌道生成のために、眼球 EYE_Y , EYE_P 及び、まぶたの関節 EYELID_P 、前頭部の皮膚を上下させる EYEBROW_P を制御する。ただし現状の HRP-4C では、左右の眼球を個別に制御できず、眼球 EYE_Y 及び EYE_P は、左右同時にヨー角とピッチ角を制御する。同様に、EYELID_P も左右のまぶたを同時に上下させる。

まず眼球 EYE_Y と EYE_P について、瞳検出（**4.1.2**）の式 (5) で得られた瞳の位置に基づいて眼球の方位角を求め、関節軸 EYE_Y の角度を決定した。具体的には、p_x を図 6 の A-C の線分間の距離で正規化して、±45 [deg] の範囲に割り当てた。ここで、眼球の上下動を制御する EYE_P に関しては常に 0 とし、EYEBROW_P についても常に 0 を与えた。

続いて、まぶたの開度は式（10）によって推定した連続値を目標とする。EYELID_P は、首動作の制御同様、モータ制御に合わせ、5ms の時間分解能にリサンプリングしてローパスフィルタとスケーリングを施した。ここで、ローパスフィルタのカットオフ周波数は EYELID_P （まばたき）に 30Hz、EYE_Y（視線）に 2Hz を用いた。

4.2.3 唇動作の生成

唇動作は、図 11 の 4 つの関節 (MOUTH_P, UPPERLIP_P, LOWERLIP_P, CHEEK_P)

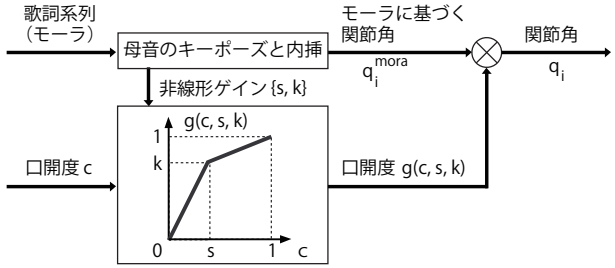

図 12 口開度に基づく唇動作の修正

によって制御される。それぞれの関節の可動範囲を**表 1** に示す。ここで、事前に行った実験では、単純に人間の口開度 c（図 10(b)）をパラメータとして与えたのでは、適切な顔動作を生成できず、それぞれの母音らしく見えなかった。これは、HRP-4C の顔内部の機構の制限が原因であり、単純に真似るだけでは、適切な動作生成が行えないことを意味する。

そのような問題を解決するために、日本語の 5 母音（/a/, /i/, /u/, /e/, /o/）と撥音（/N/）、ブレスに対応する関節角度（キーポーズ）を、それぞれの母音らしく見えるように予め定めておき（**表 2**）[27]、VocaListener で得られた歌詞とブレスの時刻情報をもとに関節軌道を生成する（**図 13**(a)）。ただし、このようなキーポーズによるパターン生成のみでは、正しいタイミングで推定された母音とブレスの唇形状だけが反映され、子音部における口の開きや、推定時刻にわずかなずれがあった場合、ゆっくりもしくは早く口を開く場合などに、それらを表現できずロバストでない。

そこでさらに、画像から得られた口開度情報 c（図 10(b)）を重畳することによって、母音やブレスの口の開き方を細かに再現し、子音に対応する動きを再現する。ここで、c はこれまで同様カットオフ周波数 20 Hz のローパスフィルタを施した。このようにして多くの場合、自然な唇軌道が生成できていることを確認した。

しかし、いくつかの音素（/i/, /u/及び/o/）において、観測される口開度が実際のキーポーズよりも小さいことがあった。これは、口開度がそれぞれのキーポーズに正規化されているわけではなく、口の開きの最大値によって正規化されていることによる。例えば、人間の/i/における口開度が 0.6 を超えることはほとんどない。したがって、修正された唇軌道は常にキーポーズの 60%以下の値となってしまう。

このような問題を解決するために非線形ゲインを導入する（**図 12**）。与えられた口開度 c とパラメータ $\{s, k\}$ から、変形のための非線形ゲイン $g(c, s, k)$ を、次式によって決定する。

Vol.2012-MUS-94 No.6
Vol.2012-SLP-90 No.6
2012/2/3

表 1 関節の動きと唇形状の関係

関節名	目的	可動範囲 (deg)
MOUTH_P	あごの開閉	0 – 10
UPPERLIP_P	上唇の上下動	-25 – 0
LOWERLIP_P	下唇の前進・後退	0 – 25
CHEEK_P	口角の上下動	-3.3 – 0

表 2 母音とブレスに関するキーポーズ

母音	/a/	/i/	/u/	/e/	/o/	/N/	ブレス
MOUTH_P [deg]	9	0	0	6	8	0	10
UPPERLIP_P [deg]	-5	-25	-23	0	-10	0	0
LOWERLIP_P [deg]	5	25	24	0	10	0	0
CHEEK_P [deg]	0	-2	0	-1	0	-1	0
非線形ゲイン s	0.5	0.3	0.3	0.6	0.6	0.5	0.5
非線形ゲイン k	0.5	0.7	0.6	0.8	0.8	0.5	0.5

$$q_i = g(c,s,k)q_i^{mora} \tag{11}$$

$$g(c,s,k) := \begin{cases} (k/s)c & (0 \le c < s) \\ \frac{1-k}{1-s}(c-s)+k & (s \le c \le 1) \end{cases}$$

ここで、q_i と q_i^{mora} は、それぞれ i 番目の口関節角とモーラに基づく関節角である。パラメータ $\{s,k\}$ は、各母音ごとに表 2 に示すように決定した。これらはモーラ系列によって変化しながら、キュービック・スプラインによって滑らかに内挿される（図 13(b)）。

図 13(c) に、実際に生成された 4 つの関節軌道と対応する唇の形状を示す。

5. 人間の歌唱に基づく歌声合成システム VocaListener のブレスを真似る歌声合成への拡張

人間の歌手は歌唱中にブレス（吸気）するため、その顔動作を真似るロボットも同様に口を開ける (**4.1.3** を参照)。しかし、口が開くのみで何も音が聞こえないと不自然な印象を与えるため、ブレス音も真似て歌声合成できるように VocaListener を拡張した。

5.1 ブレス検出手法

本稿では、我々が以前開発した、人間の歌唱中のブレスを自動検出する手法[28),29)] を用いる。ここで、ブレス/歌声/無音の 3 種の HMM (Hidden Markov Model) を構築して歌唱音声中のブレスを検出する。HMM は、RWC 研究用音楽データ（ポピュラー音楽）[23)] の

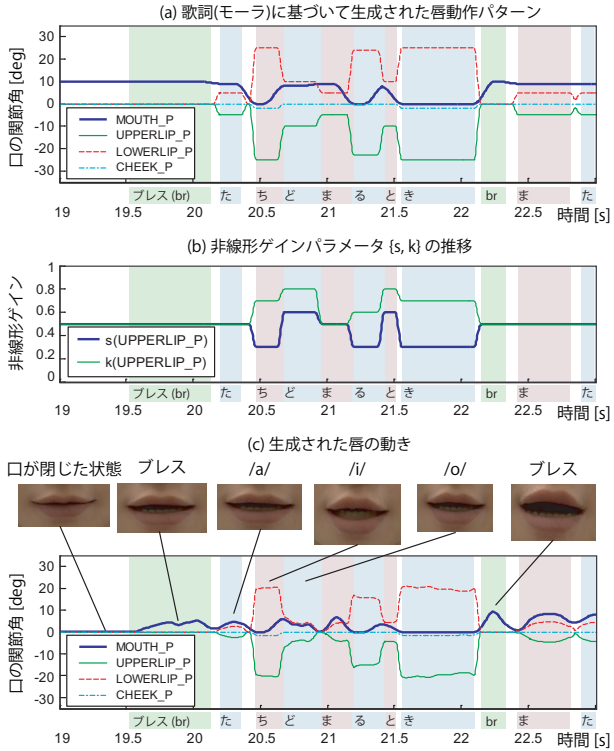

図 13 VocaListener で得られた歌詞（モーラ）と時刻情報、及び口開度 c(図 10) に基づく関節軌道の生成。(a) 歌詞に基づく口開度と (b) それらの非線形ゲイン、(c) 図 12 の処理に基づく最終的な口開度と唇形状。

27 曲と AIST ハミングデータベース[30)] 中の二人の歌唱データに手作業でラベル付けして構築した。より詳細な分析条件や楽曲名等は文献 29) で述べられている。

本手法は高い再現率を持つ一方で、呼気部や/h/等の一部の子音に対して誤検出を伴う。そこで、次のような単純な後処理によって、ブレス検出の精度を向上させる。

- 歌唱フレーズの直前以外の場所に存在する（VocaListener で推定された歌詞時刻や、歌唱フレーズの直後）検出結果を削除する
- 継続時間長が 50 ms〜1225ms の範囲[28),29)] 外の検出結果を削除する

それでも残った誤りは手作業で修正する。

Vol.2012-MUS-94 No.6
Vol.2012-SLP-90 No.6
2012/2/3

5.2 ブレス合成手法の課題

ブレス音を対象として「ユーザ歌唱を真似る」ためには、既存の VocaListener と同様、既存の歌声合成システムでブレス音を合成し、その音量パラメータをユーザ歌唱に合わせて自動的に推定する方法が考えられる。しかし、この方法は実用性・汎用性が低いため採用しない。なぜなら、音高や音量と異なり、ブレスに関するパラメータは歌声合成システムによって異なってしまう可能性が高く（場合によっては存在せず）、そのパラメータによって変化する音響的特徴がシステム毎に異なることが考えられるためである。

実際、ヤマハ株式会社の Vocaloid と Vocaloid2[5] ではブレスの合成結果が異なり、Vocaloid2 では 5 種類のブレス音を継続時間長を変えながら合成できるのに対し、Vocaloid では 1 種類のみが合成できるだけで、継続時間長も変更できない（変更しても、不適切なノイズしか合成できない）。また、Vocaloid2 でも、5 種類中のいくつかはブレスとして不自然な音であった。したがって、これまで通りの方法では、異なる歌声合成システムにおいて適用できない可能性があり、汎用的でない。

5.3 ブレス合成手法

本研究では、**5.2** 節で述べた課題を解決するために、ソースフィルタ分析に基づくブレス音合成手法を開発して、人間のブレスを真似て歌声合成する。まず合成対象のブレス音を、同じ歌声合成システム（例えば、Vocaloid2「初音ミク」）で合成する。その際、特にブレスらしい音のみを選択して用いる[⋆1]。続いて、そのブレス音のスペクトル包絡を時系列として推定し、それをブレスの時間・周波数テンプレートとして用いる。その際、本稿では、スペクトル包絡の推定に TANDEM-STRAIGHT[31] を用いた。

次に、ブレス検出（**5.1** 節）によって得られたブレス音の継続時間長と、その音量を真似るように、テンプレートを伸縮・変形させる。継続時間長は、各周波数ビンを時間方向に線形伸縮させて反映した。音量は、スペクトル包絡の周波数軸方向の積分で近似し、それを目標に合わせて変調させる。最後に、そのスペクトル包絡からインパルス応答波形を生成し、励振音源としてのガウス雑音を畳み込むことでブレス音を合成する。

このような手法を用いる事で、部分的にでもブレスらしいテンプレートが手に入れば、音量と継続時間長を変えて汎用的にブレスを合成できる。また、ブレス音が存在しない歌声合成システムでも、ブレス音合成できる可能性[⋆2] があるが、それは今後の研究課題である。

⋆1 初音ミクの場合は br5 を用いた。Vocaloid でも、部分的にブレスらしく聞こえる音を切り出して利用できる。

⋆2 ブレスの第 1, 第 2 フォルマント周波数は母音の/a/や/e/のフォルマント周波数に近い[29] という知見があり、また主観的な印象では次の歌詞の母音に応じてブレス音が変動することから、母音のスペクトル包絡そのままに、ガウス雑音による励振を行う事で近似できる。実際に試したが、場合によってはそれらしく聞こえる事もあった。しかし、ノイジーな印象が強く、包絡の変形等の処理が必要と考えられる。

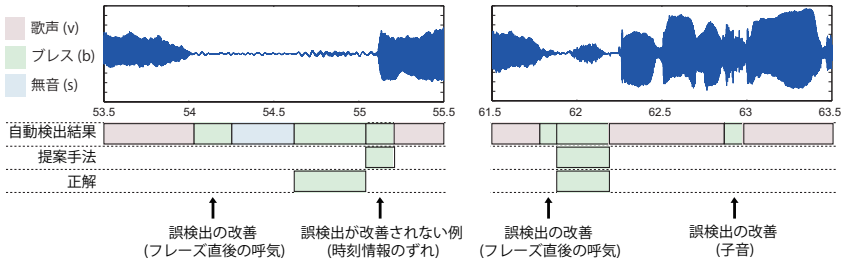

図 14 ブレス検出結果の例。フレーズ終わりの呼気や子音による誤検出が改善されたが、時刻情報がずれるなど、誤検出が改善されない場合もあった。

5.4 ブレス検出結果

実験に用いた歌唱（PROLOGUE, 約 298 秒）では、自動検出の結果、歌声/ブレス/無音区間が 289 箇所得られ、歌声区間が 152 箇所（169.71 秒）、ブレス区間が 80 箇所（20.06 秒）、無音区間が 57 箇所（109.2 秒）であった。ただし、ブレス区間の正解は 53 箇所であり、上記の初期出力結果は誤検出を含む。すなわち、再現率が 100%（= 53/53）、子音やフレーズ終わりの呼気部等で誤検出があり、精度は 66.25%であった（= 53/80）。

そこで、80 箇所から **5.1** 節の規則によって候補削除を行ったところ、53 箇所のブレス区間（17.16 秒）に絞られ、ブレス位置の検出としては再現率と精度ともに 100%であった。**図 14** にブレス検出例を示す。図 14 左に示すように、時刻のずれが残ったまま、誤検出が改善されない例が見られ、その 1 箇所のみ時刻を手で修正した。

ここでは良い結果が得られたが、歌い手や歌唱スタイルの違いによっては、ブレスの有声化や、母音の無声化によって著しく精度が下がる場合があった。今後は上述の時刻ずれの補正や、母音の無声化、ブレスの有声か等への対処に研究の余地がある。

6. 結　　果

図 15 に歌い手の女性（左）と、VocaWatcher によって生成した HRP-4C の表情（右）の比較を示す。人間に近い顔動作の生成ができたが、以下のような問題点も残った。

ロボットの口開度が人間に比べて小さい（図 15(a), (c)） これ以上口を開くことができない、ロボット関節の可動限界が原因である。

Vol.2012-MUS-94 No.6
Vol.2012-SLP-90 No.6
2012/2/3

図 15 オリジナルの人間の歌い手（左）と提案する手法によって顔動作を制御した HRP-4C（右）

人間と違いロボットの眼が閉じきっていない（図 15(b)） 過電流とモーター燃焼の問題を回避するために、口を完全に閉じきらずに少し開いた設定にしていることが原因である。

/o/, /u/の口が表現できない（図 15(b)） /o/や/u/のような口をすぼめる表情は、そのようなモーターが存在しないために表現できない。

以上の問題はすべて、今後、顔制御機構の性能向上に伴って改善される可能性がある。

また、我々のシステムの特長として、間奏のような何も歌っていない箇所でも、人間がするように、頭部を揺らしたり、視線を動かしたりといった表現を行うことができる（図 15(d)）。そういった無意識の表現も真似ることが、より自然で人間らしい動きにつながる示唆を得た。

7. おわりに

本研究では、人間に近い外観で表情制御が可能なヒューマノイドロボット HRP-4C[4]（図 1）に、歌声合成システム VocaListener を組み合わせた上で、人間の顔表情を真似て歌うための顔動作生成システム VocaWatcher を新たに実現した。また、その際にブレス音を合成できるよう VocaListener を拡張した。本研究は、最先端のロボット技術、音楽情報処理技術、画像処理技術の融合が新たな価値を生み出すことを示す意義を持つ。また、本研究の長期的な展望としては、「人間らしさ」とは何かを解明し、より人間を知ることも目指している。本成果は、人間のような歌声や動作を再現性高く人工的に生成できることから、実験での統制がとりやすい利点があり、人間の歌唱機能の解明に向けた基本ツールとして貢献できる。

本成果の実機デモンストレーションを、エンターテインメント分野における可能性を知る意味も込めて、技術展示会 CEATEC JAPAN 2010（2010 年 9 月に幕張メッセで開催）に出展した。その際、顔以外に腕も動かしたが、動作生成ソフトウェア Choreonoid[32] を用いて、手作業で音楽に合うように振り付けた。多数の来場者が訪れ、様々な反響[⋆1] が得られた。人間らしさや自然さが優れている点を高く評価する意見が多かったが、顔の動作や声の質、皮膚や顔形状などの見た目に関して、一部不自然さが残るため、気味の悪さを感じる聴衆もいた。本デモンストレーションの動画は、ウェブサイト (http://staff.aist.go.jp/t.nakano/VocaWatcher/index-j.html) で閲覧できる。

こうしたエンターテイメント分野への応用には、様々な可能性がある。歌声合成システムや歌うヒューマノイドロボットは、人間の機能を人工的に再現するだけでなく、人間の限界を超える表現[⋆2]や、クリエータが自分単独ではできない表現[⋆3]に応用可能である。表現者が人間でなくシステムやロボットであれば、クリエータの立場からは、気兼ねすることなく、自分のイメージする世界を柔軟な発想でそのまま表出できる利点がある。同じ声質でも様々なクリエータが違った歌い方や世界観を表現したり、同じヒューマノイドロボットでも違った表情を見せたりすることで、表現がより多様になる可能性がある。また、そのように一つのロボットやシステムが多様な表現を持っていれば、リスナーの立場からは、複数のロボットやシステムから好みのものを選んで、それぞれの中から好みの表現を選択して楽しむこともできる。さらに、ロボットやシステムが歌うことによる驚きと楽しさが加えられるだけでなく、ロボットやシステムが歌うからこそ意味があったり感動できる歌詞など、新たな楽しみの創出に繋がる可能性がある。

今後の課題として、ロボット関節の軌道生成には、いくつかのゲインパラメータや事前に設定するパラメータが含まれており、それらは HRP-4C に特化してしまっている。VocaListener で歌声合成の音源の違いを吸収する上で反復推定が効果的であったように、今後、VocaWatcher でも同様の発想で反復推定を導入していくことで、様々なヒューマノイドロボットに対応できる予定である。本研究では、「模倣」を出発点として「自然さ」をまずは表現することが重要だと考えたが、次の段階として、そのモデル化（コンテキストの時

⋆1 例えば、http://www.diginfo.tv/2010/10/13/10-0217-r-en.php や http://blogs.wsj.com/japanrealtime/2010/10/05/japans-next-pop-idol-is-a-robot/ 。
⋆2 高い歌や早い歌を歌う、同じ動きで歌う等。
⋆3 男性クリエータが女性の歌声・振付でコンテンツを作る等。

間変化とパラメータ空間内での制御点の時間変化の対応関係の機械学習）に関する研究を進めることで、模倣を越えた新たな表現へつなげていきたいと考えている。

謝辞 本研究では、ヤマハ株式会社及び、クリプトン・フューチャー・メディア株式会社の歌声合成ソフトウェア「初音ミク（CV01）」、RWC 研究用音楽データベース (ポピュラー音楽 RWC-MDB-P-2001) 及び AIST ハミングデータベースを使用した。本研究を推進するに当たって三浦 加奈子 氏、米倉 健太 氏、松本 吉央 氏、比留川 博久 氏、関口 智嗣 氏、からサポートを得た。

参 考 文 献

[1] 濱崎雅弘，武田英明，西村拓一：動画共有サイトにおける大規模な協調的創造活動の創発のネットワーク分析–ニコニコ動画における初音ミク動画コミュニティを対象として–，人工知能学会論文誌，Vol. 25, No. 1, pp. 157–167 (2010).

[2] 濱野 智史：インターネット関連産業，デジタルコンテンツ白書 2009，pp. 118–124 (2009).

[3] Kenmochi, H.: VOCALOID and Hatsune Miku phenomenon in Japan, *Proc. of InterSinging 2010*, pp. 1–4 (2010).

[4] Kaneko, K., Kanehiro, F., Morisawa, M., Miura, K., Nakaoka, S. and Kajita, S.: Cybernetic Human HRP-4C, *Proc. of Humanoids 2009*, pp. 7–14 (2009).

[5] Kenmochi, H. and Ohshita, H.: VOCALOID – Commercial Singing Synthesizer based on Sample Concatenation, *Proc. of Interspeech 2007*, pp. 4010–4011 (2007).

[6] Tachibana, M., Nakaoka, S. and Kenmochi, H.: A Singing Robot Realized by a Collaboration of VOCALOID and Cybernetic Human HRP-4C, *Proc. of InterSinging 2010*, pp. 9–14 (2010).

[7] Nakano, T. and Goto, M.: VocaListener: A Singing-to-Singing Synthesis System Based on Iterative Parameter Estimation, *Proc. SMC 2009*, pp. 343–348 (2009).

[8] 中野倫靖，後藤真孝：VocaListener：ユーザ歌唱の音高および音量を真似る歌声合成システム，情報処理学会論文誌，Vol. 52, No. 12, pp. 3853–3867 (2011).

[9] Kajita, S., Nakano, T., Goto, M., Matsusaka, Y., Nakaoka, S. and Yokoi, K.: VocaWatcher: Natural Singing Motion Generator for a Humanoid Robot, *Proc. of IROS 2011*, pp. 2000–2007 (2011).

[10] 梶田秀司，中野倫靖，後藤真孝，松坂要佐，中岡慎一郎，横井一仁：ヒューマノイドロボットの自然な歌唱動作生成，第 29 回日本ロボット学会学術講演会，pp. 1–4 (2011).

[11] Kato, I., Ohteru, S., Shirai, K., Matsushima, T., Narita, S., Sugano, S., Kobayashi, T. and Fujisawa, E.: The Robot Musician WABOT-2 (Waseda robot-2), *Robotics*, Vol. 3, pp. 143–155 (1987).

[12] Chida, K., Okuma, I., Isoda, S., Saisu, Y., Wakamatsu, K., Nishikawa, K., Solis, J., Takanobu, H. and Takanishi, A.: Development of a New Anthropomorphic Flutist Robot WF-4, *Proc. of ICRA 2004*, pp. 152–157 (2004).

[13] Mizumoto, T., Tsujino, H., Takahashi, T., Ogata, T. and Okuno, H.: Thereminist Robot: Development of a Robot Theremin Player with Feedforward and Feedback Arm Control based on a Theremin's Pitch Model, *Proc. of IROS 2009*, pp. 2297–2302 (2009).

[14] Sawada, H., Nakamura, M. and Higashimoto, T.: Mechanical voice system and its singing performance, *Proc. of IROS 2004*, Vol. 2, pp. 1920–1925 (2004).

[15] Kuroki, Y., Fujita, M., Ishida, T., Nagasaka, K. and Yamaguchi, J.: A Small Biped Entertainment Robot Exploring Attractive Applications, *Proc. of ICRA 2003*, pp. 471–476 (2003).

[16] Murata, K., Nakadai, K., Yoshii, K., Takeda, R., Torii, T., Okuno, H. G., Hasegawa, Y. and Tsujino, H.: A Robot Singer with Music Recognition Based on Real-time Beat Tracking, *Proc. of ISMIR 2008*, pp. 199–204 (2008).

[17] Lina, C.-Y., Chenga, L.-C., Tsenga, C.-K., Gub, H.-Y., Chungb, K.-L., Fahnb, C.-S., Lub, K.-J. and Changc, C.-C.: A Face Robot for Autonomous Simplified Musical Notation Reading and Singing, *Robotics and Autonomous Systems*, Vol. 59, pp. 943–953 (2011).

[18] Wilbers, F., Ishi, C. and Ishiguro, H.: A Blendshape Model for Mapping Facial Motions to an Android, *Proc. of IROS 2007*, pp. 542–547 (2007).

[19] Jaeckel, P., Campbell, N. and Melhuish, C.: Facial Behavior Mapping — From Video Footage to a Robot Head, *Robotics and Autonomous Systems*, Vol. 56, pp. 1042–1049 (2008).

[20] 後藤真孝，齋藤 毅，中野倫靖，藤原弘将：歌声情報処理の最近の研究，日本音響学会誌，Vol. 64, No. 10, pp. 616–623 (2008).

[21] Goto, M., Saitou, T., Nakano, T. and Fujihara, H.: Singing Information Processing Based on Singing Voice Modeling, *Proc. of ICASSP 2010*, pp. 5506–5509 (2010).

[22] Janer, J., Bonada, J. and Blaauw, M.: Performance-driven control for sample-based singing voice synthesis, *Proc. 9th Int. Conference on Digital Audio Effects (DAFx-06)*, pp. 41–44 (2006).

[23] 後藤真孝，橋口博樹，西村拓一，岡 隆一：RWC 研究用音楽データベース:研究目的で利用可能な著作権処理済み楽曲・楽器音データベース，情報処理学会論文誌，Vol. 45, No. 3, pp. 728–738 (2004).

[24] Matsumoto, Y., Ino, T. and Ogasawara, T.: Development of Intelligent Wheelchair System with Face and Gaze Based Interface, *Proc. of ROMAN 2001*, pp. 262–267 (2001).

[25] Morris, T., Blenkhorn, P. and Zaidi, F.: Blink Detection for Real-time Eye Tracking, *Journal of Network and Computer Applications*, Vol. 25, pp. 129–143 (2002).

[26] Seeing Machines: `http://www.seeingmachines.com/`.

[27] Nakaoka, S., Kanehiro, F., Miura, K., Morisawa, M., Fujiwara, K., Kaneko, K., Kajita, S. and Hirukawa, H.: Creating Facial Motions of Cybernetic Human HRP-4C, *Proc. of Humanoids 2009*, pp. 561–567 (2009).

[28] Nakano, T., Ogata, J., Goto, M. and Hiraga, Y.: Analysis and automatic detection of breath sounds in unaccompanied singing voice, *Proc. 10th International Conference of Music Perception and Cognition (ICMPC 10)* (2008).

[29] 中野倫靖，後藤真孝， 緒方淳， 平賀譲：ブレスの合図を認識する伴奏システムの実装と評価，情報処理学会研究報告音楽情報科学 2008-MUS-76，Vol. 2008, No. 50, pp. 83–88 (2008).

[30] 後藤真孝，西村拓一：AIST ハミングデータベース：歌声研究用データベース，情報処理学会研究報告，2005-MUS-61, pp. 7–12 (2005).

[31] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. and Banno, H.: TandemSTRAIGHT: A Temporally Stable Power Spectral Representation for Periodic Signals and Applications to Interference-free Spectrum, F0, and Aperiodicity Estimation, *Proc. of ICASSP 2008*, pp. 3933–3936 (2008).

[32] Nakaoka, S., Kajita, S. and Yokoi, K.: Intuitive and Flexible User Interface for Creating Whole Body Motions of Biped Humanoid Robots, *Proc. of IROS 2010*, pp. 1675–1682 (2010).