情報処理学会論文誌 Vol.55 No.3 1222–1235 (Mar. 2014)

VocaWatcher：ユーザ歌唱の顔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,a)] 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]

受付日 2013年3月19日, 採録日 2013年12月4日

概要：本論文では，ユーザ歌唱における顔表情を真似てヒューマノイドロボットの顔動作を生成するVocaWatcherについて述べる．ここで，我々が以前開発したVocaListenerを用い，ユーザ歌唱の歌い方（音高と音量）を真似て歌声合成も行う．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，ロボットによるより自然な歌唱を実現するために，我々が以前開発したブレス音の検出技術とVocaListenerを組み合わせ，ブレス音を真似て合成できるように拡張した．

キーワード：VocaListener，VocaWatcher，歌声合成，ヒューマノイドロボット，顔動作生成

VocaWatcher: A Facial-motion Generation System for Humanoid Robot by Imitating Facial Expressions of User's Singing

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

Received: March 19, 2013, Accepted: December 4, 2013

Abstract: In this paper, we describe *VocaWatcher* that is a facial-motion generator for a singing robot by imitating user's singing. It can synthesize singing voices by using our previous VocaListener to imitate pitch (F0) and dynamics (power) of user's singing. Although singing humanoid robots have been developed with synthesized singing voices, such robots do not appear to be natural because of limitations of manual control. To generate natural singing expressions, VocaWatcher imitates a human singer by analyzing a video clip of 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 by combining our previous method of breath sound detection to imitate breathing sounds that make the robot singing more realistic.

Keywords: VocaListener, VocaWatcher, singing synthesis, humanoid robot, facial-motion generation

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，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，歌声合成技術およ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（コンテンツ分野）」[*1]における「アイドルロボット」構想に見られるように，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テインメン

[1] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305–8568, Japan
[a)] vocawatcher-ml@aist.go.jp

[*1] http://www.meti.go.jp/policy/economy/gijutsu_kakushin/kenkyu_kaihatu/str2010download.html

© 2014 Information Processing Society of Japan

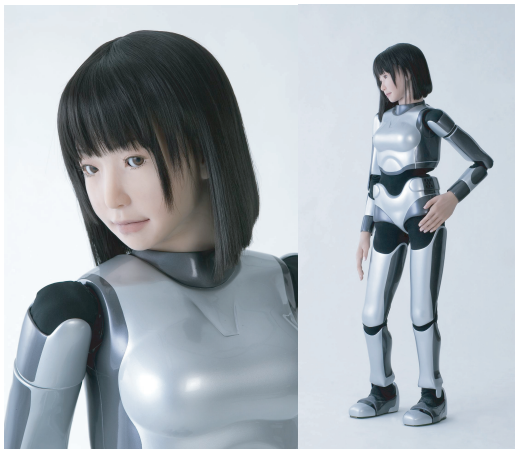


図 1 ヒューマノイドロボット HRP-4C の外観
Fig. 1 Appearance of a humanoid robot HRP-4C.

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 1 つに歌声合成がある．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 [1] 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 [2] を開発してきた（図 **1**）．具体的には，身長 160 cm，体重 46 kg（バッテリ含む），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．HRP-4C においては，ヤマハ株式会社と共同で，歌声合成ソフトウェア VOCALOID2 [3] を用いて歌唱させる展示も行うなど [4]，ヒューマノイドロボット技術のエンタ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 [5]（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** 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ともにブレス（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．

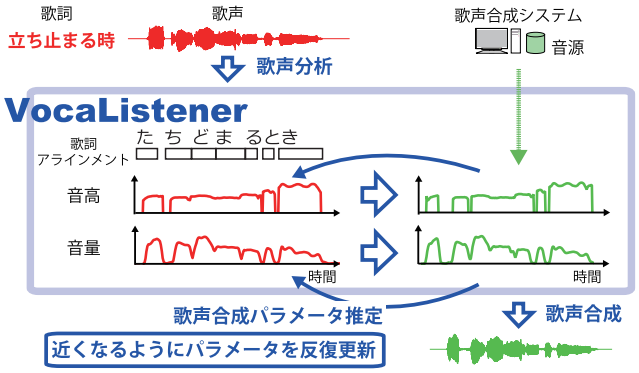


図 2 VocaListener の処理概要．人間の歌声と歌詞を入力として，その音高（F_0）と音量に近くなるように歌声合成パラメータを反復推定して歌声合成する
Fig. 2 Overview of VocaListener that takes a singing voice of a human singer and its lyrics as the input and synthesizes a singing voice by iteratively estimating singing synthesis parameters to imitate singer's pitch (F_0) and dynamics (power).

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 [6] から始まり，フルート演奏 [7]，テルミン演奏 [8] などが存在する．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 [9]，アカペラ歌唱やダンス可能なロボット [10]，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 [11]，簡略化された楽譜映像を認識して歌う顔ロボット [12] などが研究されてきた．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りした．

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て，モーションキャプチャ結果 [13] や，人のビデオ映像の顔追跡結果 [14] を入力として用いる研究がある．しかしこれらは，顔へのマーカ付与が必要であったり [13]，特定個人ごとの顔モデル（Active Appearance Models: AAMs）の学習を必要としたりする [14] 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 [15] 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，

© 2014 Information Processing Society of Japan

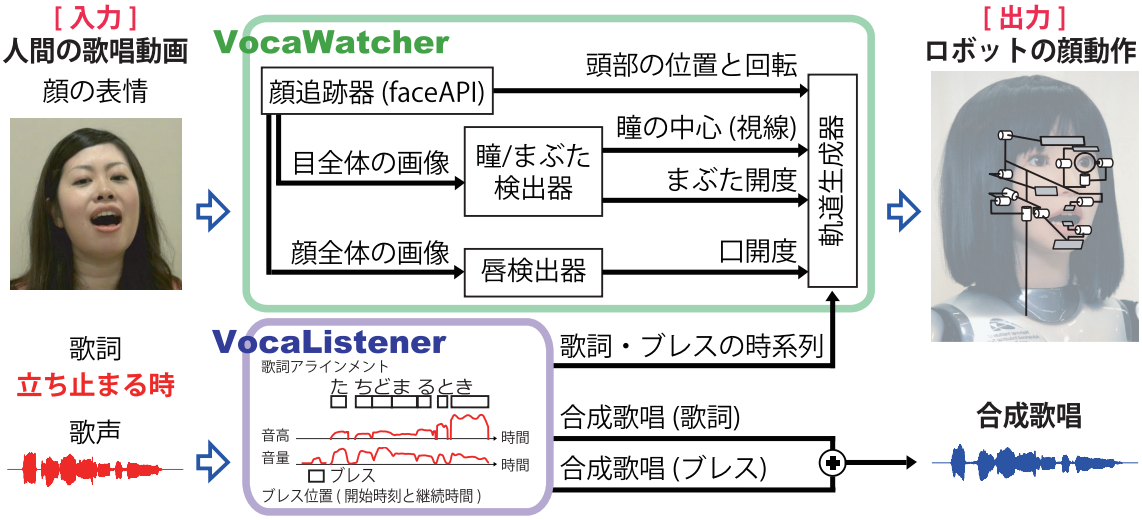


図 3 VocaWatcher による人間の歌唱を真似るヒューマノイドロボットの顔動作生成の処理概要

Fig. 3 Overview of VocaWatcher that generates facial expressions of a humanoid robot by imitating a human singer.

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（たとえば VOCALOID [3]）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）[5]．パラメータの反復推定により，推定精度が従来研究 [16] 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[*2]．独自の歌声専用音響モデルによって，歌詞のテキストを歌詞を音符ごとに割り当てる作業は，ほぼ自動で行える．ここで，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．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の時刻情報を入力として与え，「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，各音節の開始時刻は，母音（/t a/であれば/a/部分）の開始時刻が出力される．また，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 1,920 × 1,080（29.97 fps）で収録したが，VocaWatcher では，その解像度をすべて 960 × 540 にリサンプリングして使用した．ここでは，映像の適用可能性を高めるために，普及している機材の性能や設定を考慮して，解像度のリサンプリングを行い，一般的によく用いられているフレームレートを設定した．これによって，新たに録画して模倣する利用方法だけでなく，既存の録画を模倣することも視野に入れている．また今回は，顔動作の表現範囲として，単一のカメラで表現できる範囲を対象とした．つまり，カメラの視野から外れたり，顔が隠れてしまったりするほど歌い手が動くことなどは想定していない．

図 4 歌唱収録風景．右端の女性歌手を左端のカメラで撮影し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

Fig. 4 A recording scene with the target human singing. We used a video clip in which a female singer on the right was recorded using a video camera on the left and a singing voice recorded by using a microphone.

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）[17] の「PROLOGUE」（RWC-MDB-P-2001 No.7，298.2 秒）を使用して，日本人女性 1 名による歌唱を収録した．歌声合成システムとしては

[*2] 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．

© 2014 Information Processing Society of Japan

「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

人間の歌い手の顔表情分析は，本節で述べる動画像処理によって行う．まず歌い手の顔を追跡して頭部の位置と回転や，目，鼻，口の位置を推定する．その後，これらの推定結果を用いて瞳とまぶたの検出，口開度の検出を行う．

ここで従来，瞳検出に関する研究は，視線検出に基づく車いす制御 [18] や，まばたき検出に基づくコマンド制御 [19] 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が歌唱中の人間は歌唱動作としてつね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とらえる必要がある．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，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．

本節では，顔表情分析の詳細として，以上の問題を解決する手法についても説明する．これらはすべて，不特定の歌い手の表情を分析することを想定して実装した．

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．本論文では，Seeing Machine 社の顔画像トラッキングソフトウェア faceAPI*4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．

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．たとえば，

*3 http://www.crypton.co.jp/mp/pages/prod/vocaloid/cv01.jsp
*4 http://www.seeingmachines.com/

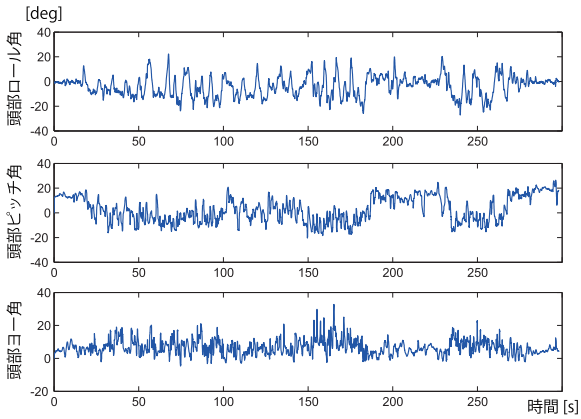


図 5 歌唱動画から推定された歌い手の頭部の回転
Fig. 5 Singer's head rotation estimated from the video clip of a singing performance.

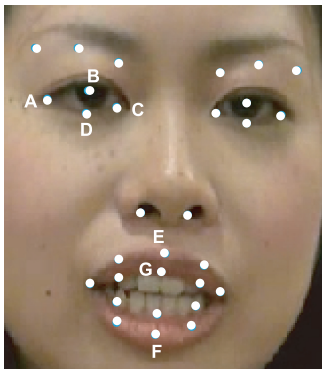


図 6 歌手の顔から検出された特徴点
Fig. 6 Feature points detected from singer's face.

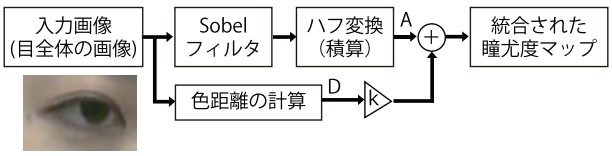


図 7 色による重みを加えたハフ変換に基づく瞳検出の概要
Fig. 7 Overview of iris detection based on the Hough transform weighted by color distance.

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．

4.1.2.1 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 としたときに，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

© 2014 Information Processing Society of Japan

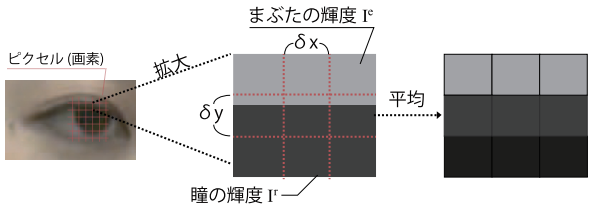


図 8 サブピクセル情報を用いたまぶた検出のための分解能向上におけるピクセルと実画像の関係

Fig. 8 Relation between pixel and real image in resolution improvement by using the subpixel information for eyelid detection.

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．

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

本論文では，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 1 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．

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

4.1.2.2 サブピクセル情報による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 3〜6 [pixel] と少なかった．このように，通常のピクセルベースの検出では，まぶた開度の推定に少ない離散値しか用いることができず，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目領域の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，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．

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

ただし，現在の実装では，4.1.2.1 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 1 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．

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

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

ここで，e は瞳全体を大きなピクセルと見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．

© 2014 Information Processing Society of Japan

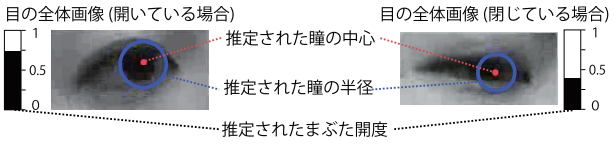


図 9 瞳の中心と半径，まぶた開度の推定結果の例

Fig. 9 An example of estimating the center position and radius of the iris, and the aperture rate.

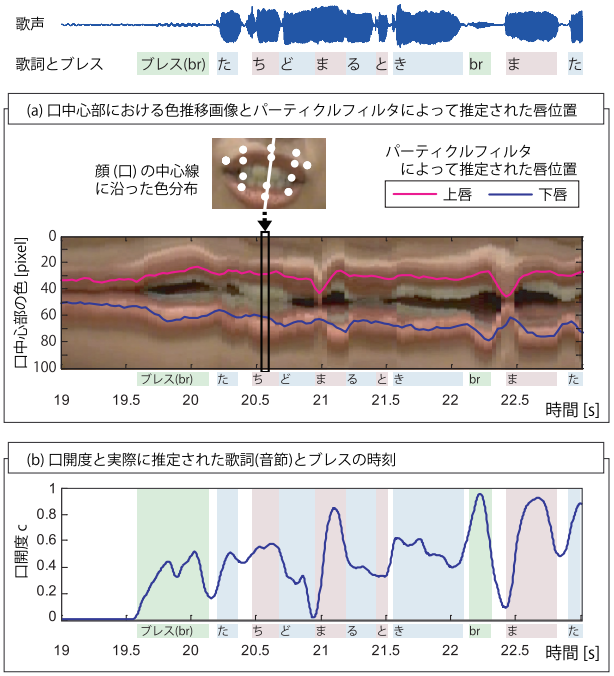


図 10 (a) パーティクルフィルタによって推定された唇の動き，(b) 口開度と VocaListener によって推定された歌詞とブレスの時刻の関係

Fig. 10 (a) Lip motion estimated by the particle filter and (b) relation between the mouth aperture ratio and temporal alignment of lyrics and breath estimated by VocaListener.

4.1.3 口開度の検出

口開度は，顔追跡（4.1.1 項）によって得られた唇の特徴点から推定する．しかし，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持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 (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．

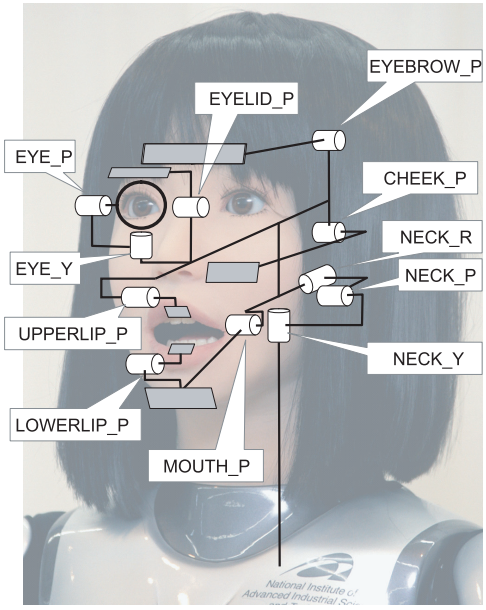


図 11 HRP-4C の顔と首の関節軸構成 [20]．円柱がモータ，平行四辺形が皮膚を変形させるための機構の動作端，右目の円は眼球を示す．それぞれの関節名の末尾で，制御可能な回転軸を示しており，「_R(ロール軸)」「_P(ピッチ軸)」「_Y（ヨー軸)」である

Fig. 11 The face and neck mechanism of HRP-4C [20]. The cylinders and the parallelograms indicate motors and the actuation points to deform the skin, respectively. The circle on the right eye is an eyeball. The joint name appendices "_R", "_P", and "_Y" denote roll, pitch, and yaw axes, respectively.

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頭部の位置と回転，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 [20]．ここで，それぞれの関節角をサーボモータにより 5 ms 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，29.97 fps 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，5 ms の時間分解能に線形補間してリサンプリングするが，その際には，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

© 2014 Information Processing Society of Japan

表 1 唇動作に関係する関節の可動範囲
Table 1 Movable range of each joint related to lip motion.

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

EYE_Y，EYE_P および，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 および 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 項）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 に関してはつねに 0 とし，EYEBROW_P についてもつねに 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 30 Hz，EYE_Y（視線）に 2 Hz 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）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，それぞれの唇形状（母音など）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，それぞれの唇形状らしく見えるようにあらかじめ定めておき（**表 2**）[20]，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 12** (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撥音・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，ゆっくり口を開いたり速く口を開いたりする場合などに，それらを表現できずロバストでない．

表 2 母音とブレスに関するキーポーズ
Table 2 Key poses for vowels and breath.

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

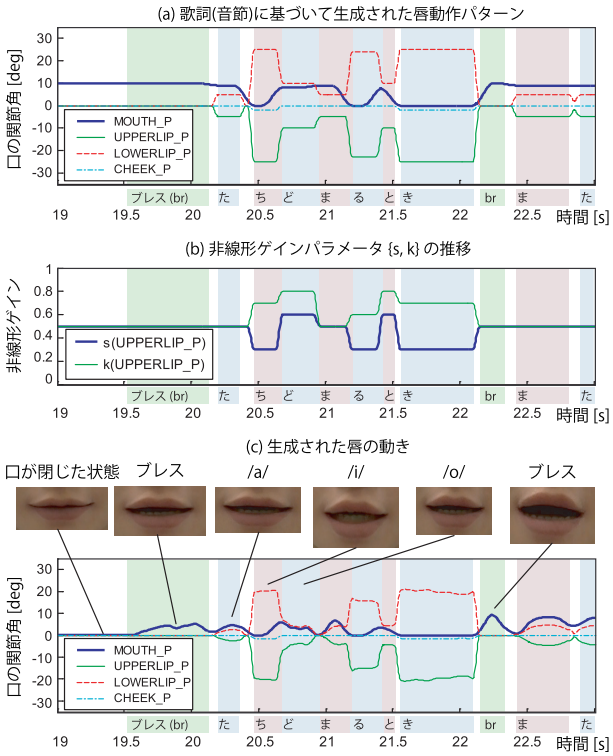


図 12 口開度 c（図 10）に基づく関節軌道の生成．(a) 歌詞（母音）に基づく口開度と (b) それらの非線形ゲイン，(c) 図 13 の処理に基づく最終的な口開度と唇形状
Fig. 12 Pattern generation based on the mouth aperture ratio c (Fig. 10). (a) Mouth aperture ratio based on vowel types in the lyrics, (b) corresponding nonlinear gain parameters, and (c) mouth aperture and lip shapes generated after the adjustment of Fig. 13.

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/および/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，口開度が口の開きの最大値によって正規化されていることが原因である．たとえば，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つねにキーポーズの 60%以下

© 2014 Information Processing Society of Japan

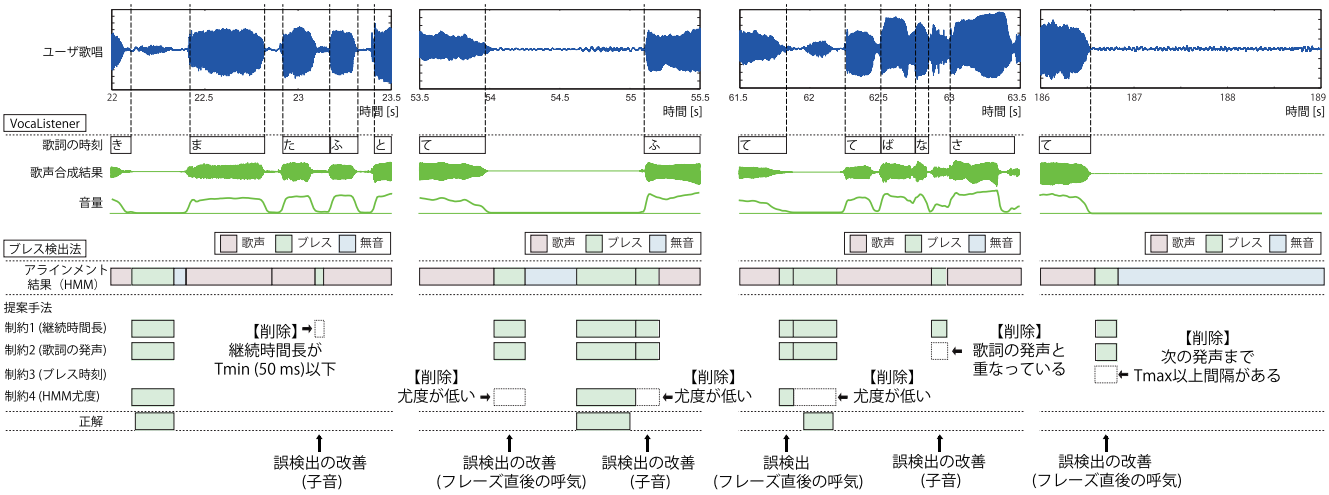


図 14 ブレス検出結果と誤検出を改善するための追加処理（制約による候補削除）

Fig. 14 A result of the breath detection and additional processing to eliminate false detections (candidate elimination based on constraints).

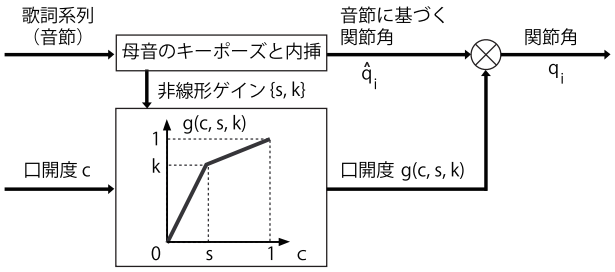


図 13 推定された人間の口開度に基づく唇動作の修正

Fig. 13 Lip motion modification based on the estimated mouth aperture ratio of the human singer.

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 13**）．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．

$$q_i = g(c,s,k)\hat{q}_i \tag{11}$$

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

ここで，q_i と $\hat{q}_i$ は，それぞれ i 番目の口関節角と音節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．これらは音節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 12 (b)）．

図 12 (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 項を参照）．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．

本章では，ブレス検出手法とブレス合成手法について述べるが，特にブレス検出結果は，4.2.3 項における唇動作生成においてブレスに対応する関節角度（キーポーズ）を決定するうえでも必要となる（表 2）．

5.1 ブレス検出手法

本論文では，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 [21], [22] 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 の構築には，特徴量として音声認識で広く用いられている MFCC（Mel-Frequency Cepstrum Coefficient），ΔMFCC，ΔPower を利用し，それぞれ 3 状態 HMM でモデル化し，各状態の分布は対角共分散行列を用いた 16 混合ガウス分布とした．HMM は，RWC 研究用音楽データ（ポピュラー音楽）[17] の 27 曲と AIST ハミングデータベース [23] 中の 2 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などは文献 [22] で述べられている．

ただし，本手法は不特定歌唱者の歌声に対して高い再現率を持つ一方で，呼気部や/h/などの一部の子音に対して誤検出をともなう．そこで，ブレスの最小継続時間を $T_{\min}$ ms，最大継続時間を $T_{\max}$ ms として以下のような後処理を今回新たに導入し，ブレス検出の精度を向上させる（**図 14**）．

制約 1（継続時間長） 継続時間長が $[T_{\min}, T_{\max}]$ の範囲外の検出結果を削除する．

制約 2（歌詞の発声） 子音による誤検出を減らすため，

© 2014 Information Processing Society of Japan

VocaListener で合成された歌詞の発声と重なる（区間の始端から終端までの音量がすべて 0 より大きい）検出結果を削除する．

制約 3・制約 4（フレーズ終端の呼気に関する 2 つの制約） 歌唱フレーズ終端の呼気による誤検出を減らすため，歌唱フレーズ直前に存在する検出結果を 1 つ選ぶ．

制約 3（ブレス時刻） VocaListener で合成された歌詞の発声（直前が無音）の先頭時刻から $T_{\max}$ 前までの区間以外の検出結果を削除する．

制約 4（HMM 尤度） VocaListener で合成された歌詞の発声（直前が無音）の先頭時刻から $T_{\max}$ 前までの区間に複数の候補があった場合，ブレス HMM の尤度が最も高かった検出結果を 1 つ選択し，それ以外を削除する．

ここで，$T_{\min} = 50\,\text{ms}$，$T_{\max} = 1{,}225\,\text{ms}$ とした [21], [22]．それでも残った誤りは手作業で修正する．以上の制約は，制約 1〜4 まで順番に（縦続に）適用した．ここで，制約 1 と制約 2 の順序は入れ替えても問題ないが，これらは制約 3 以前に行われることを想定している．

5.2 ブレス検出結果

今回，実験に用いた歌唱は，3 章で説明したように，RWC 研究用音楽データベース（ポピュラー音楽）[17] の「PROLOGUE」（RWC-MDB-P-2001 No.7，298.2 秒）を歌った日本人女性 1 名によるものである．この歌い手および歌声は，ブレス検出用の HMM を学習させるためのデータ [21], [22] には含まれていない．本歌唱においては，手作業で付与したブレス区間の正解ラベルが 53 カ所であり，自動検出の結果，歌声/ブレス/無音すべての区間として 289 カ所得られ，そのうち歌声区間が 152 カ所（169.71 秒），ブレス区間が 80 カ所（20.06 秒），無音区間が 57 カ所（109.2 秒）であった．ここで，提案手法の有効性を評価するために，ブレス検出の再現率（R）と精度（P）を算出した．R と P は，それぞれ以下のように定義する．

$$R = \frac{\text{events correctly detected as breath}}{\text{events in breath}} \times 100 \quad (12)$$

$$P = \frac{\text{events correctly detected as breath}}{\text{events detected as breath}} \times 100 \quad (13)$$

ここで，検出結果が正解のブレス位置と時間的に重なりがあれば正解とした．$R = 100\%$（$= 53/53$）であったのに対し，子音やフレーズ終わりの呼気部などでの誤検出によって $P = 66.25\%$（$= 53/80$）であった．

そのようにして自動推定された 80 カ所のブレス検出箇所の候補について，5.1 節の制約 1（継続時間長）によって 80 → 77 へ，制約 2（歌詞の発声）によって 77 → 72 へ，制約 3（ブレス時刻）によって 72 → 71 へ，制約 4（HMM 尤度）によって 71 → 53 へ，候補が削除された．図 14 にブレス検出例を示す．誤った候補削除が行われて，時間的に近傍に存在した呼気（0.1 秒のずれ）が選択された例が 1 カ所あり（図 14：右から 2 番目に示した箇所），その時刻のみを手で修正した．このようにして得られた 53 カ所のブレス区間は，1 カ所の誤検出によって再現率が低下したが，検出精度は改善して $R = 98.11\%$（$= 52/53$），$P = 98.11\%$（$= 52/53$）となった．

ここで，HMM によるブレス検出手法 [21], [22] による初期検出結果において，歌い手や歌唱スタイルの違いによってブレスが有声化する（音高を感じるブレス）場合，それを歌声として検出してしまい，再現率が低下することがあった．本論文で述べた制約は，初期検出結果の再現率が高い場合には有効だが，そもそもブレスとして検出できなかった場合への対処が難しい．今後は，そのようなブレス音を追加して HMM を再学習させたり適応させたりするか，異なる特徴量を検討するなど，ブレスの有声化への対処に研究の余地がある．また，今回提案した 4 つの制約も改善の余地がある．たとえば，前述したように，長い間奏の間には息を吸う必要があるため，曲と歌手によってはフレーズ間でも口を開いて息を吸う可能性があるが，そのようなブレスは今回の方法では削除されてしまう．そのほか，今回の制約を HMM のネットワーク（状態遷移）として組み入れたり，顔から検出された特徴点など（図 6）を活用したりする拡張も考えられる．

5.3 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ごと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 [3] 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，これまでどおりの方法では，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．

5.4 ブレス合成手法

本研究では，5.3 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

© 2014 Information Processing Society of Japan

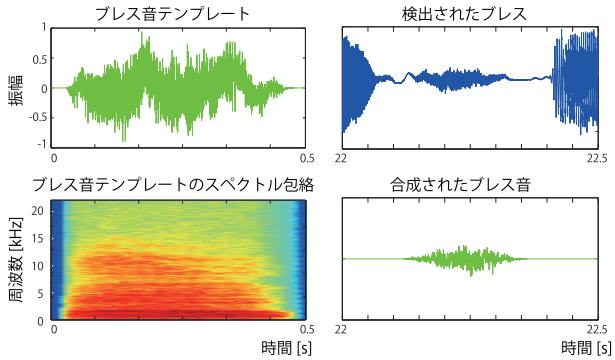


図 15 ブレス音テンプレートの波形とスペクトル包絡の時間変化の例（左），およびユーザ歌唱のブレスとそのテンプレートからの合成結果の例（右）

Fig. 15 Examples of the waveform of a breath template and its spectral envelope (on the left), and a user's breath sound and the corresponding synthesized result (on the right).

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で合成する．その際，聴感上ブレスらしい音と感じた音のみを選択して用いる．続いて，そのブレス音のスペクトル包絡の時系列を推定し，それをブレスの時間・周波数テンプレートとして用いる（**図 15**）．本論文では，歌声合成システムとして用いた「VOCALOID2 初音ミク」において「br5」と指定することで合成できるブレス音をテンプレートとして用いた．また，スペクトル包絡の推定には TANDEM-STRAIGHT [24] 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．そのため，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，このようなスペクトル包絡をガウス雑音で励振させる手法は，ブレス音が存在しない歌声合成システムからも，ブレス音を生成できる可能性がある．たとえば，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 [22] という知見があるため，それらの母音のスペクトル包絡をガウス雑音による励振を行うことで近似できる可能性がある．また，主観的な印象ではあるが，次から始まる歌詞の母音に応じて，そのブレスのブレス音が変動する場合があるため，その母音のスペクトル包絡を同様に用いることも考えられる．これらの方法を実際に試したが，場合によってはそれらしく聞こえることもあった．ただし，上記の方法そのままでは，合成された音にノイジーな印象が強い場合が多く，包絡の変形などの何らかの処理が必要と考えられ，今後の研究課題である．

6. 結果と考察

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）を真似て歌った結果は，http://staff.aist.go.jp/t.nakano/VocaWatcher/index-j.html で閲覧できる．これまで説明してきた「PROLOGUE」（RWC-MDB-P-2001 No.7，298.2 秒）を用いた結果に加え，同一の日本人女性 1 名と異なる日本人女性 1 名による歌唱を真似て，歌声合成システム「VOCALOID2 メグッポイド」と「VOCALOID2 VY1」で，それぞれ新たな別の 2 曲について，顔動作の生成と歌声合成を行った結果のデモンストレーション動画がある．

図 16 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 16 (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．

人間と違いロボットの眼が閉じきっていない（図 16 (b)）
過電流とモータ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 16 (b)） /o/ や /u/ のような口をすぼめる表情は，そのようなモータが存在しないために表現できない．

以上の問題が原因で，歌詞の読み取り（リップリーディング）が困難になることがあるが，これらは今後，顔制御機構の性能向上にともな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 16 (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という示唆を得た．

最後に，生成されたロボットの関節軌道（制御パラメータ）として，技術展示会 CEATEC2009（2009 年 10 月に幕張メッセで開催）に出展した際の制御パラメータ [4] と VocaWatcher による制御パラメータの生成結果を，頭部の回転と口開度について**図 17** と**図 18** にそれぞれ示す．生成結果を比較すると，頭部の回転については，従来手法の結果が人工的な軌跡を描いていることが分かる（図 17）．さらに，従来は動作ポイントを（曲を聴きながら）手動で指定する必要がある点 [4] も，VocaWatcher と異なる．また口開度については，VocaWatcher による生成では，同じ母音でもフレーズ中の位置に応じて値が異なっている（図 18．たとえば，母音/a/では「た ちどま るとき また ふとふりかえる」）．また，ブレス（br）位置も適切に付与

© 2014 Information Processing Society of Japan

(a) 歌い出し，/ta/ を発声
(20.29 秒，デモ動画中の時刻は 00:35:51)

(c) /ra/ を発声
(151.48 秒，デモ動画中の時刻は 02:46:68)

(b) 目を閉じた状況，/ru/ を発声
(74.17 秒，デモ動画中の時刻は 01:29:39)

(d) 間奏中，発声なし
(204.77 秒，デモ動画中の時刻は 03:39:99)

図 16 人間の歌い手の顔（左）と提案手法によって顔動作を生成した HRP-4C の顔（右）
Fig. 16 Examples of the face of a human singer (on the left) and the face of HRP-4C (on the right) whose facial motions were generated by the proposed method.

されており，後続の口開度に応じた値の変化や，ブレスの継続時間長の（必要に応じた）違いなどがあった．これらが，自然性や表現力の向上に寄与したと考えられる．

ただし，たとえ正しい関節軌道（制御パラメータ）が再現されても，ロボットによる自然な歌唱動作として最適である保証はできない．なぜなら，生成される歌唱動作の自然さは，ロボットの見た目や性能（ハードウェアの制約など）にも依存するなど，複合的な要素に基づいて決定されると考えられるためである．本論文では人間を模倣した制御パラメータを生成することで自然性の向上を目指したが，より自然な歌唱動作の生成へ向けて，今後も研究を進めていきたい．

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 [2]（図 1）に，歌声合成システム VocaListener を組み合わせたうえで，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．従来の，手作業で個々の動作を指定するインタフェースに加え，VocaWatcher のように顔の動きで入力できるインタフェースによって，ロボット制御ツールの多様性を増してユーザの多様な要求に応えやすい．また本研究は，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．

実機デモンストレーションの結果と考察

本成果の実機デモンストレーションを，エンタ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 [25] 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[*5]が得られた．顔動作（VocaWatcher）や歌声やブレス音（VocaListener）について，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，6 章で示したような唇動作の問題点など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．ただし，歌唱動作の不自然さや気味の悪さを感じる原因としては，皮膚の材質や顔形状などの見た目など，本論文の範囲を超える原因も含まれていると考えられる．したがって，歌唱動作の自然性を考える際，場合によっては，そのような評価軸の考慮が必要だと考えられる．

[*5] http://staff.aist.go.jp/t.nakano/VocaWatcher/index-j.html にメディア報道，記事などの一覧を示す．

© 2014 Information Processing Society of Japan

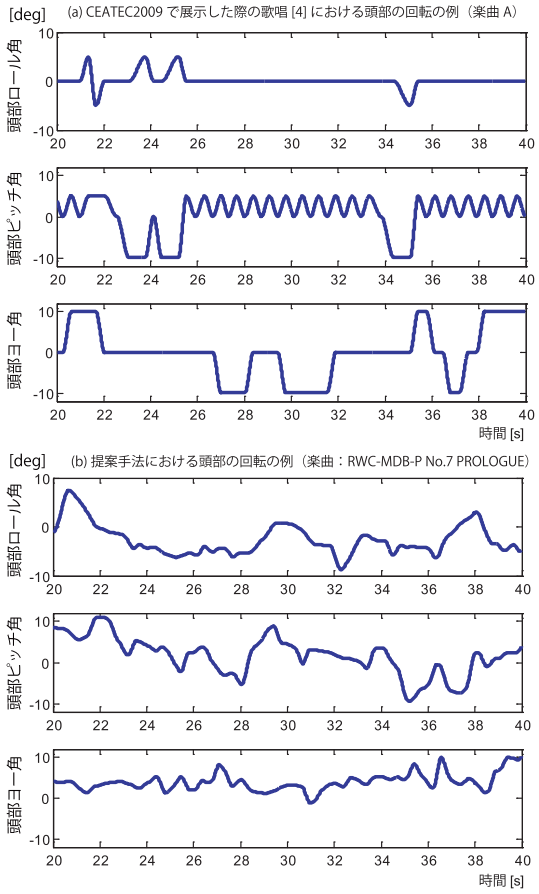


図 17 (a) 従来手法 [4] と (b) VocaWatcher による頭部の回転
Fig. 17 Results of head rotation of (a) a previous approach and (b) VocaWatcher.

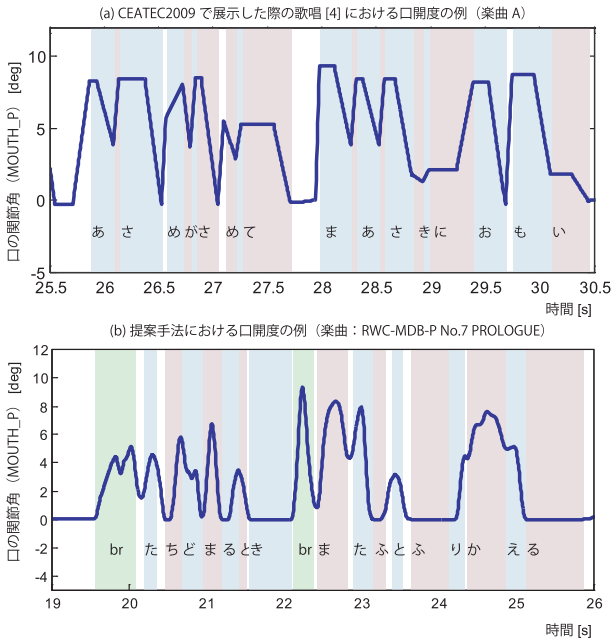


図 18 (a) 従来手法 [4] と (b) VocaWatcher による口開度
Fig. 18 Results of mouth aperture of (a) a previous approach and (b) VocaWatcher.

こうしたエンタテインメント分野への応用には，様々な可能性がある．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や，クリエータが自分単独ではできない表現をするために応用することができる．たとえば，前者では同じ歌唱動作を再現性高く行うなどであり，後者では男性クリエータが女性の歌声と振付でコンテンツを作るなどである．表現者が人間でなくロボットであれば，クリエータの立場からは，自分のイメージする世界を柔軟な発想で気兼ねなく表現できるという利点がある．同じロボットでもクリエータによって違った歌い方や表情を見せることで，表現がより多様になる可能性がある．また，リスナの立場からは，好みのロボットの好みの歌唱表現を選択して楽しむなど，選択の自由度が増えるという利点がある．さらに，ロボット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，ロボットが歌うからこそ意味があったり感動できたりする歌詞など，新たな楽しみの創出につながる可能性がある．

今後の展開

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うえで反復推定が効果的であった [5] ように，今後，VocaWatcher でも同様の発想で反復推定を導入するなど，様々なヒューマノイドロボットへ対応できる手法の実現を目指す．また，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て，対象となる楽曲を実際に歌っている人間の映像があることを前提とした手法を提案した．次の段階として，そのモデル化（コンテキストの時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，任意の楽曲への対応など，模倣を超えた新たな表現へつなげられる可能性がある．

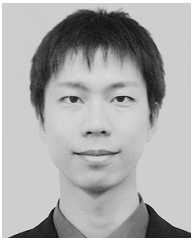
謝辞 本研究では，歌声合成ソフトウェア「VOCALOID2 初音ミク」，「VOCALOID2 メグッポイド」，「VOCALOID2 VY1」を使用した．また，RWC 研究用音楽データベース（ポピュラー音楽 RWC-MDB-P-2001）および AIST ハミングデータベースを使用した．本研究を推進するに当たって，三浦加奈子氏，米倉健太氏，松本吉央氏，比留川博久氏，関口智嗣氏，からサポートを得た．

参考文献

[1] 後藤真孝：初音ミク，ニコニコ動画，ピアプロが切り拓いた CGM 現象，情報処理（情報処理学会誌），Vol.53, No.5, pp.466–471 (2012).
[2] Kaneko, K., Kanehiro, F., Morisawa, M., Miura, K., Nakaoka, S. and Kajita, S.: Cybernetic Human HRP-4C, *Proc. 9th IEEE-RAS International Conference on Humanoid Robots* (*Humanoids 2009*), pp.7–14 (2009).

© 2014 Information Processing Society of Japan

[3] Kenmochi, H.: VOCALOID and Hatsune Miku Phenomenon in Japan, *Proc. 1st Interdisciplinary Workshop on Singing Voice* (*InterSinging 2010*), pp.1–4 (2010).
[4] Tachibana, M., Nakaoka, S. and Kenmochi, H.: A Singing Robot Realized by a Collaboration of VOCALOID and Cybernetic Human HRP-4C, *Proc. 1st Interdisciplinary Workshop on Singing Voice* (*InterSinging 2010*), pp.9–14 (2010).
[5] 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).
[6] 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).
[7] 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. 2004 IEEE International Conference on Robotics and Automation* (*ICRA 2004*), pp.152–157 (2004).
[8] 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. 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS 2009*), pp.2297–2302 (2009).
[9] Sawada, H., Nakamura, M. and Higashimoto, T.: Mechanical voice system and its singing performance, *Proc. 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS 2004*), Vol.2, pp.1920–1925 (2004).
[10] Kuroki, Y., Fujita, M., Ishida, T., Nagasaka, K. and Yamaguchi, J.: A Small Biped Entertainment Robot Exploring Attractive Applications, *Proc. 2003 IEEE International Conference on Robotics and Automation* (*ICRA 2003*), pp.471–476 (2003).
[11] 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. 9th International Conference on Music Information Retrieval* (*ISMIR 2008*), pp.199–204 (2008).
[12] 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).
[13] Wilbers, F., Ishi, C. and Ishiguro, H.: A Blendshape Model for Mapping Facial Motions to an Android, *Proc. 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS 2007*), pp.542–547 (2007).
[14] 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).
[15] 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).
[16] Janer, J., Bonada, J. and Blaauw, M.: Performance-driven Control for Sample-based Singing Voice Synthesis, *Proc. 9th International Conference on Digital Audio Effects* (*DAFx-06*), pp.41–44 (2006).
[17] 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).
[18] Matsumoto, Y., Ino, T. and Ogasawara, T.: Development of Intelligent Wheelchair System with Face and Gaze Based Interface, *Proc. 10th IEEE International Workshop on Robot and Human Interactive Communication* (*ROMAN 2001*), pp.262–267 (2001).
[19] 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).
[20] 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. 9th IEEE-RAS International Conference on Humanoid Robots* (*Humanoids 2009*), pp.561–567 (2009).
[21] 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).
[22] 中野倫靖，後藤真孝，緒方　淳，平賀　譲：無伴奏歌唱におけるブレスの音響特性とそれに基づく自動ブレス検出，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).
[23] 後藤真孝，西村拓一：AIST ハミングデータベース：歌声研究用音楽データベース，情報処理学会研究報告音楽情報科学研究会，2005-MUS-61, pp.7–12 (2005).
[24] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. and Banno, H.: TANDEM-STRAIGHT: A Temporally Stable Power Spectral Representation for Periodic Signals and Applications to Interference-free Spectrum, F0, and Aperiodicity Estimation, *Proc. 2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP 2008*), pp.3933–3936 (2008).
[25] Nakaoka, S., Kajita, S. and Yokoi, K.: Intuitive and Flexible User Interface for Creating Whole Body Motions of Biped Humanoid Robots, *Proc. 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS 2010*), pp.1675–1682 (2010).



中野 倫靖 （正会員）

2003 年図書館情報大学卒業．2008 年筑波大学大学院図書館情報メディア研究科博士後期課程修了．博士（情報学）．現在，産業技術総合研究所主任研究員．日本音響学会会員．2006 年日本音楽知覚認知学会研究選奨，2007 年インタラクション 2007 インタラクティブ発表賞，2009 年情報処理学会山下記念研究賞（音楽情報科学研究会），2010 年音楽情報科学研究会（夏のシンポジウム 2010）ベストプレゼンテーション賞各受賞．

© 2014 Information Processing Society of Japan

後藤 真孝 （正会員）

1998 年早稲田大学大学院理工学研究科博士後期課程修了．博士（工学）．同年電子技術総合研究所に入所し，2001 年に改組された産業技術総合研究所において，現在，情報技術研究部門首席研究員兼メディアインタラクション研究グループ長．IPA 未踏 IT 人材発掘・育成事業プロジェクトマネージャー等を兼任．ドコモ・モバイル・サイエンス賞基礎科学部門優秀賞，科学技術分野の文部科学大臣表彰若手科学者賞，情報処理学会長尾真記念特別賞，星雲賞等，35 件受賞．

梶田 秀司

1985 年東京工業大学大学院修士課程修了（制御工学専攻）．同年通産省工業技術院機械技術研究所に入所．2 足歩行ロボット等の動的制御技術の研究に従事．1996 年 2 月より 1 年間米国カリフォルニア工科大学客員研究員．2001 年より組織改変にともない独立行政法人産業技術総合研究所主任研究員，現在に至る．1996 年 3 月東京工業大学より学位取得（工学博士）．著書『歩き出した未来の機械たち』（ポプラ社），『ヒューマノイドロボット』（編著）（オーム社）．1996 年度計測自動制御学会論文賞，2005 年度日本ロボット学会論文賞各受賞．

松坂 要佐

2003 年早稲田大学大学院理工学研究科修了．博士（工学）．2003 年日本学術振興会特別研究員 PD，2004 年より 2005 年まで南カリフォルニア大学客員研究員を経て，2006 年産業技術総合研究所特別研究員，2008 年同研究所研究員として画像認識およびマルチモーダルシステムに関する研究に従事．現在，（株）MID アカデミックプロモーションズ役員．

中岡 慎一郎

2006 年東京大学大学院情報理工学系研究科コンピュータ科学専攻博士課程修了．博士（情報理工学）．同年産業技術総合研究所に入所．2012 年より知能システム研究部門主任研究員，現在に至る．2011 年 11 月より 1 年間英国エジンバラ大学客員研究員．ヒューマノイドロボットによる動き提示，ロボットソフトウェアプラットフォーム等の研究に従事．日本ロボット学会会員．2008 年度日本ロボット学会論文賞，IEEE/SICE SII2012 Best Paper Award（Robotics）各受賞．

横井 一仁

1986 年東京工業大学大学院機械物理工学専攻修了．同年工業技術院機械技術研究所に入所．2001 年産業技術総合研究所知能システム研究部門主任研究員．2009 年同所同部門副研究部門長，現在に至る．1995～1996 年スタンフォード大学客員研究員，2005 年より筑波大学大学院教授（連携大学院），2013 年より技術研究組合国際廃炉研究開発機構兼務．人間型ロボット等の研究に従事．2005 年日本ロボット学会論文賞，2008 年日本機械学会賞（論文）ほか受賞．日本機械学会フェロー，日本ロボット学会フェロー，IEEE 等の会員．博士（工学）．

© 2014 Information Processing Society of Japan