ヒューマノイドロボットの自然な歌唱動作生成

○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 (産総研)

1. はじめに

本稿では、歌声に合わせたリアルなロボットの顔動作を生成する技術、VocaWatcher について述べる。ヒューマノイドロボットに歌を歌わせる試みとしては、Kuroki *et al.*[1] や、Murata *et al.*[2] が存在するが、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。2009 年、我々は平均的な日本人青年女性のプロポーションと表情制御可能なヒューマノイドロボット HRP-4C を開発した [3]。2010 年にはこのロボットを用いて、ヤマハ株式会社との共同研究により歌唱合成ソフトウェア Vocaloid2 を用いて歌を歌わせる実験を行っている [4]。ここでは、顔動作の自動生成と人手によるプログラミングが併用されたが、歌い手としての表現力には限界があった。本研究では、人間の歌い手の表情と歌声をもとに動作生成を行うことによって、エンターテインメントロボットに要求されるより幅の広い表現力をロボットに与えることを目指している。

2. 人間の歌唱の収録

対象とする楽曲として RWC 研究用音楽データベース（ポピュラー音楽）[7] で提供される「PROLOGUE」(RWC-MDB-P-2001 No.7) を使用した。収録の様子を図 1 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て以降の処理を進めた。

図 1 歌唱収録風景

3. 人間の歌唱に基づく歌声合成

3·1 VocaListener: 人間の歌唱に基づく歌声合成パラメータの自動推定

VocaListener [8] は、既存の歌声合成ソフトウェア (例えば Yamaha の Vocaloid [9]) 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)[8]。パラメータの反復推定により、推定精度が従来研究 [10] 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。独自の歌声専用音響モデルによって歌詞のテキストを与えるだけで、そのモーラ[1] を音符毎に割り当てる作業は、ほぼ自動で行える。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]。

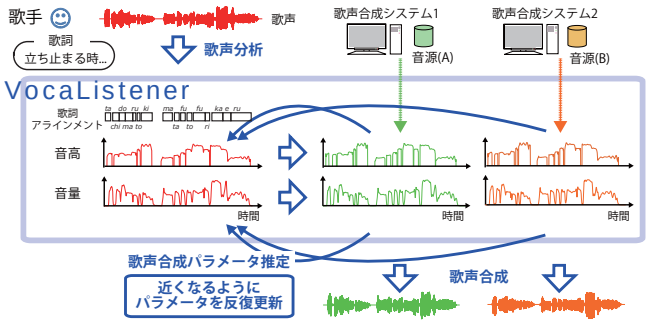

図 2 VocaListener の全体構成。人間の歌声と歌詞をもとに繰り返し計算によってパラメータ推定を行う

3·2 人間のブレスを真似る歌声合成

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てしまう (4·3 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、人間のブレスも真似て歌声合成できるように VocaListener を以下のように拡張した。

人間の歌唱中のブレス音検出については、Nakano *et al.*[11] の技術を改良し、検出精度を高めたものを使用した。合成用のブレス音に関しては、Vocaloid2 のものを利用し、*TANDEM-STRAIGHT* [12] によって、そのスペクトル包絡の時系列を推定した。さらに、ブレス検出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、それらを伸縮・変形させて合成した。

4. 顔画像解析と動作生成

図 3 に HRP-4C の頭部の関節軸構成を示す。単眼のビデオ画像と前節で得られた発音タイミングをもとにこれらの関節軸の動作パターンをいかに生成するかがここでの課題である。画像情報から、ロボットの顔動作を生成する先行研究は存在するが、被験者の顔にマーカーが必要であったり [5]、多くの学習とチューニングを要する [6] など、我々の目的に合致した手法は存在しなかった。

[1] 日本語歌唱における発音とその音符への割り当てを、3 種類のモーラ表記 “V”, “CV”, “N”（C：子音、V：母音、N：撥音）に分類して扱う。

[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。

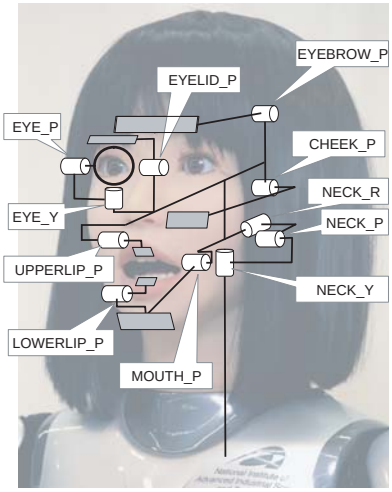

図 3 HRP-4C の顔と首の関節軸構成 [16].

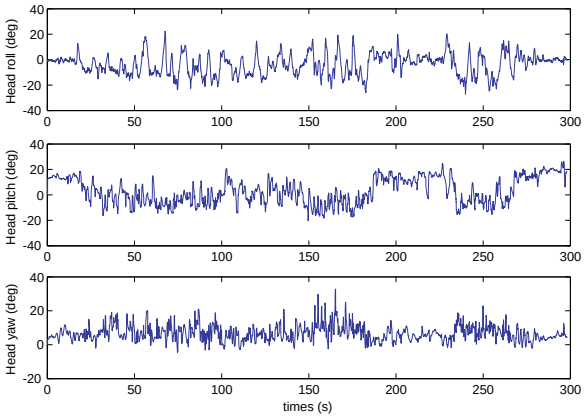

図 4 画像から推定された歌い手の頭部の動き

4·1 顔トラッキング

ビデオ映像（解像度 960×540,29.97FPS）を処理する最初のステップとして Seeing Machine 社の顔画像トラッキングソフトウェア faceAPI を用いた [13]。faceAPI はビデオ画像中の顔画像を検出し、三次元空間における頭部の位置と姿勢を推定する機能をもつ。図 4 に歌唱ビデオから検出された曲の初めから終わりまで 298.2s の間の頭部の姿勢 (roll,pitch,yaw) を示す。ロボットの首関節 NECK_R, NECK_P, NECK_Y はこれらのデータにフィルタリングとスケーリングを施して生成した。

faceAPI はビデオの各フレームにおける顔の特徴点 (Face landmark) の座標も出力する。検出された特徴点の例を図 5 に示す。点 $0, 601, 1, 602$ は右目の領域に対応する。現状の faceAPI(FaceTrackingAPI 3.2) では、瞬きを検出できないため、歌い手が右目を閉じた場合でも、点 $601, 602$ の間の距離は変化しない。この問題に対処する方法を次に述べる。

4·2 視線と瞬きの検出

視線と瞬きを画像から検出する試みとして Matsumoto *et al.*[14] や Morris *et al.*[15] がある。しかし、歌唱中の人間は、感情表現として半目を開くなど、従来の技術では検出できない動きが存在する。

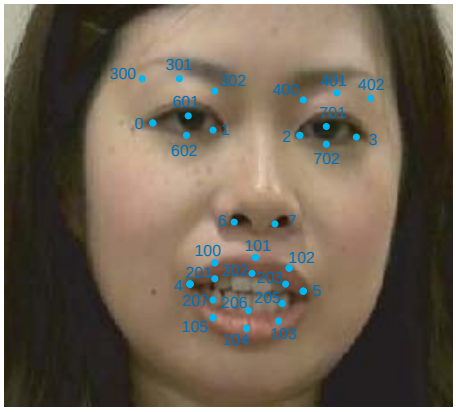

図 5 検出された特徴点 (face landmarks)．これらの点を参照して以降の画像処理を進める

視線と瞬きの検出には、faceAPI によって検出された目領域（図 5) に対して以下の処理を適用する（図 6)。

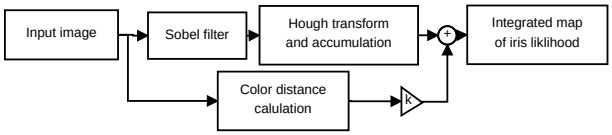

図 6 視線・瞬き検出の概要

領域内の色による重みを加えたハフ変換: 歌い手の瞳は歌唱の途中で長い時間半目の状態を保つことがあり通常の方法では安定した瞳の検出が困難であった。瞳の検出を頑健にするために、ハフ変換による瞳の検出に領域内の色の重みを加えるアルゴリズムを開発した。アルゴリズムは以下の定式を用いる：

$$L = A + kD \tag{1}$$

A は円形ハフ変換による投票結果（形を手がかりとした瞳の存在確率)、D は色距離から算出した尤度画像（色を手がかりとした瞳の存在確率)、k は重み付け定数であり、L が最終的な瞳の尤度マップである。

円形ハフ変換には円の半径 p_r が必要であるが、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。

この実験においては、歌い手が日本人であるため、色距離はモノクロ画像（瞳は黒と仮定）から算出した。

瞳の位置 p は L から以下のように決定する。

$$p = \arg\max_{xy} L_{xy} \tag{2}$$

この情報をもとに眼球の方位角を求め図 3 の関節軸 EYE_Y の角度を決定した。眼球の上下動を制御する EYE_P に関しては常に 0 としている。

サブピクセル情報を用いた瞬き検出: 今回撮影した歌唱データにおいては頭部全体を撮影する必要性から目領域の解像度は、3 から 6 ピクセルしか確保することができなかった。通常のピクセルベースのアルゴリ

ズムを使った場合は、瞳の開度は 3 から 6 の離散値でしか得ることができないが、これでは HRP-4C の歌唱表現を満足に行うことができない。より精度の高い開度を得るために、次式で表されるサブピクセル情報を用いたアルゴリズムを開発した。

$$a = \begin{cases} 0 & (e < e_{min}) \\ \frac{e-e_{min}}{e_{max}-e_{min}} & (e_{min} \le e < e_{max}) \\ 1 & (e \ge e_{max}) \end{cases} \quad (3)$$

$$e := \sum_{p_x-p_r}^{p_x+p_r} I_{xy}, \quad e_{min} := 2I^e p_r, \quad e_{max} := 2I^r p_r,$$

ここで p は式 2 で検出した瞳の位置、p_r は瞳の半径、a はまぶたの開度である。瞳の輝度 I^r は定数、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。得られた情報を元に図 3 の関節 EYELID_P の角度を決定した。

4·3 口開度の検出

ここでは口開度を上唇と下唇間の距離と定義する。これは図 5 に示した Face landmark から得られるべきものであるが、我々の実験では歌唱時の高速な唇の動きにのために faceAPI はしばしば唇のトラッキングに失敗し、正確な口開度を検出できなかった。

高速な唇の運動に対応した口開度を求めるため、我々はまず 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 101 – 104 に平行で点 202 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 7)。上下の唇は、この図の中で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、上唇の中心線 y_U および下唇の中心線 y_L を推定した。推定された唇の運動が図 7 のマゼンタと紺のラインで示されている。上下の唇の距離を $[0, 1]$ の範囲で正規化して口開度 c とした。

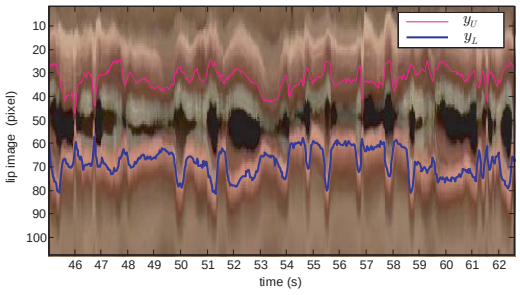

図 7 パーティクルフィルタによって推定された唇の動き

図 8 では歌い出しにおける口開度と、VocaListener で得られたモーラを比較している。発声していない期間 (19.6s-20.2s, 22.1s-22.5s) でも口が大きく開いているが、これは歌い手が息を吸いこんでいる期間（ブレス）を示している。

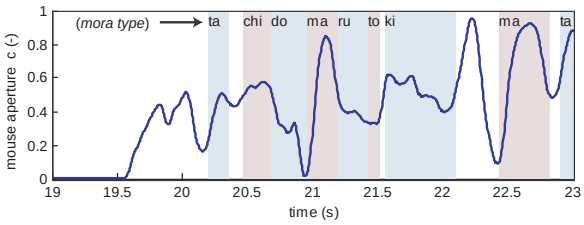

図 8 口開度とモーラの比較

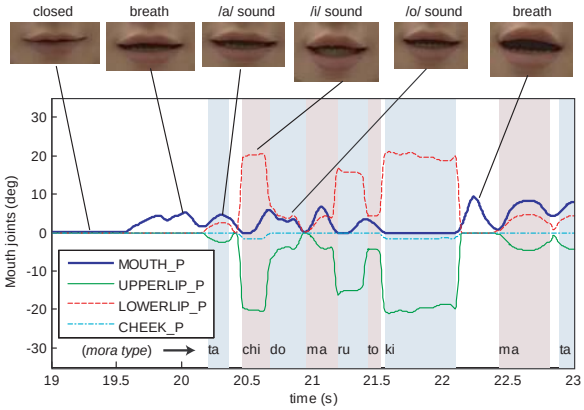

図 9 生成された唇の動き

4·4 唇動作の生成

唇動作は、図 3 の 4 つの関節 (MOUTH_P, UPPERLIP_P, LOWERLIP_P, CHEEK_P) によって制御される。日本語の 5 母音 (a,i,u,e,o) に対応する関節角度を予め定めておき [16]、VocaListener で得られたモーラの種類とタイミング情報をもとに関節軌道を生成する。これに前述の手法で画像から得られた口開度情報 c を重畳することによって、息継ぎや子音に対応する動きを再現することが可能となる。図 9 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. 結果

図 10 にオリジナルの歌い手（左）と、VocaWatcher によって生成した HRP-4C の表情（右）を同一タイミングで比較した。口の開度が十分でない、目が閉じきっていない等の問題点はあるものの、オリジナルに近い振る舞いをロボットに与えることに成功している。

2010 年 9 月に幕張メッセで開催された CEATEC JAPAN 2010 において、本技術のデモンストレーションを行い大きな注目を集めることに成功した。同じデモンストレーションはウェブサイト (http://staff.aist.go.jp/t.nakano/VocaWatcher/) で見ることができる。

謝辞

本研究を推進するに当たって協力をいただいた産業技術総合研究所ヒューマノイド研究グループの皆さん、特に三浦加奈子氏、米倉健太氏に感謝いたします。サービスロボット研究グループ長の松本吉央氏には画像処理に関して多くのアドバイスを頂きました。また、比留川博久知能システム研究部門長、関口智嗣情報システム研究部門長による寛大なるサポートに深く感謝い

(a) 歌い出し，/ta/ を発声， time = 20.29s, frame = 608

(b) 目を閉じた状況，/ru/ を発声 , time = 74.17s, frame = 2223

(c) /ra/ を発声， time = 151.48s, frame = 4540

(d) 間奏中，発声なし , time =204.77s, frame = 6137

図 10 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）

たします。

参考文献

[1] Y. Kuroki, M. Fujita, T. Ishida, K. Nagasaka and J. Yamaguchi, “A small biped entertainment robot exploring attractive applications,” in *Proc. of ICRA2003*, pp.471-476, 2003.

[2] K. Murata, K. Nakadai, *et al.*, “A robot singer with music recognition based on real-time beat tracking,” in *Proc. of ISMIR 2008 – Session 2b – Music Recognition and Visualization*, pp.199-204, 2008.

[3] K.Kaneko, F.Kanehiro, M.Morisawa, K.Miura, S.Nakaoka and S.Kajita “Cybernetic Human HRP-4C,” in *Proc. of IEEE/RSJ Int. Conference on Humanoid Robots*, pp.7-14, 2009.

[4] M.Tachibana, S.Nakaoka and H.Kenmochi, “A singing robot realized by a collaboration of VOCALOID and Cybernetic Human HRP-4C,” in *Proc. of InterSinging2010*, Tokyo, 2010.

[5] F. Wilbers, C. Ishi and H. Ishiguro, “A blendshape model for mapping facial motions to an android,” in *Proc. of the IROS2007*, pp.542-547, 2007.

[6] P. Jaeckel, N. Campbell, C. Melhuish, “Facial behavior mapping — From video footage to a robot head, ” *Robotics and Autonomous Systems*, vol.56, pp.1042-1049, 2008.

[7] M. Goto *et al.*, “RWC music database: Music genre database and musical instrument sound database,” in *Proc. of ISMIR2003*, pp.229-230, 2003.

[8] T.Nakano and M.Goto, “VocaListener: A singing-to-singing synthesis system based on iterative parameter estimation,” in *Proc. of the SMC 2009*, pp.343-348, 2009.

[9] H. Kenmochi and H. Ohshita, “Vocaloid – Commercial singing synthesizer based on sample concatenation,” in *Proc. of Interspeech 2007*, pp. 4010-4011, 2007.

[10] J. Janer, J. Bonada, and M. Blaauw, “Performance-driven control for sample-based singing voice synthesis,” in *Proc. of DAFx-06*, pp.41-44, 2006.

[11] T. Nakano, J. Ogata, M. Goto, and Y. Hiraga, “Analysis and automatic detection of breath sounds in unaccompanied singing voice,” in *Proc. of ICMPC 10*, 2008.

[12] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, “Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation,” in *Proc. of ICASSP 2008*, pp. 3933-3936, 2008.

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

[14] Y. Matsumoto, T. Ino and T. Ogasawara, “Development of intelligent wheelchair system with face and gaze based interface,” in *Proc. of ROMAN 2001*, pp.262-267, 2001.

[15] T. Morris, P. Blenkhorn and F. Zaidi, “Blink detection for real-time eye tracking,” *Journal of Network and Computer Applications*, Volume 25, Issue 2, pp.129-143, 2002.

[16] S. Nakaoka, F. Kanehiro, K. Miura, *et al.* “Creating facial motions of Cybernetic Human HRP-4C,” in *Proc. of Humanoids2009*, pp.561-567, 2009.