「音楽制作と情報処理の友好関係」特集号　解　説

人間の歌い方を真似る歌声合成システム VocaListener とロボット顔動作生成システム VocaWatcher

後藤　真孝*・中野　倫靖*・梶田　秀司*・松坂　要佐*・中岡慎一郎*・横井　一仁*

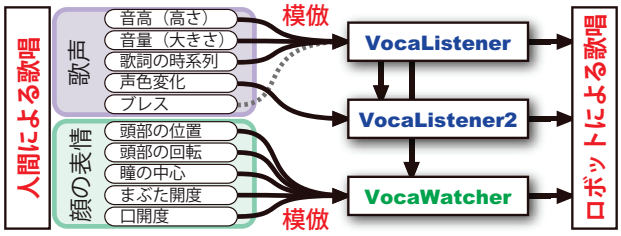
1. はじめに

歌声合成は，社会に不可欠な重要な技術になりつつある．2007年以来，VOCALOID「初音ミク」に代表される歌声合成ソフトウェアが注目を集め，歌声合成技術に基づく楽曲が動画コミュニケーションサービス「ニコニコ動画」に大量に投稿されて，大規模な協調的創造活動が起きている [1,2]．VOCALOID 関連の CD は，商業音楽ヒットチャート「オリコン」のアルバム週間ランキングで 1 位を獲得し [3],「初音ミク」の CG 映像を伴うライブコンサートも日本，アメリカ，シンガポールで開催された．「人間の歌声でなければ聴く価値がない」という旧来の価値観を打破し，合成された歌声がメインボーカルの楽曲を多くの人々が積極的に楽しむ文化が成立することが，世界で初めて実証されたのである．

しかし，人間らしい自然な歌声を誰でも自在に合成可能な技術は普及しておらず，表情豊かな歌声を合成するには，複雑で時間のかかるパラメータ調整を手作業でする必要があった．そこでわれわれは，人間の歌い方を本人とは違う声質や顔で再現するように「真似る」アプローチに取り組んだ．そして，人間らしい表情豊かな歌声をさまざまな声質で手軽に合成できる歌声合成システム VocaListener[4,5] および VocaListener2[6,7] を実現し，その後，ヒューマノイドロボットがそれと同期した自然な顔の動きで歌唱できる顔動作生成システム VocaWatcher[8,9] を実現した．

これらの歌声合成システムにより，パラメータ調整の知識が乏しい人でも自然に合成できれば，より多くの人々が歌声を合成して創作しやすくなる．しかも，歌声によってどのような表現をするのか，どのようなメッセージを伝えるのかに注力して歌声を合成できる可能性がある．さらに，人間を真似て高品質な歌声合成技術の実現を目指すことは，合成技術自体の進展に加え，人間の歌声知覚・生成機構の解明にもつながっていく．

一方，顔動作生成システムにより，ヒューマノイドロボットが自然な表情と歌声で歌うことができれば，エンターテインメント用途でより一層魅力的に活用できる．われわれの展示経験から，そもそも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，「歌うロボット」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．将来，ロボットにとって人間との自然なコミュニケーションは重要な機能であり，本研究が達成する「自然さ」を実現する技術は，声の音響信号処理技術や顔の画像処理技術，それらに基づくロボット制御技術と相まって，未来社会において大切な基礎技術になると考えている．



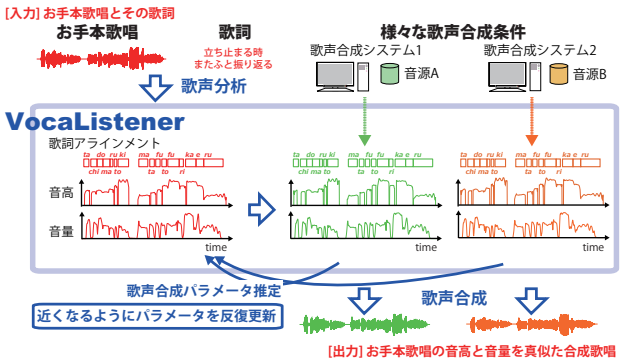
第 1 図　人間の歌い方を真似る VocaListener および VocaListener2，VocaWatcher

本解説では，人間の歌声の音高と音量を真似る歌声合成システム VocaListener[4,5]，音高と音量だけでなく声色変化も真似る歌声合成システム VocaListener2[6,7]，人間が歌唱するときの顔の表情を真似るヒューマノイドロボット用顔動作生成システム VocaWatcher[8,9] の三つの歌声情報処理システム [10,11] と，それらの関連研究を紹介する．VocaListener と VocaListener2 は，歌声のみを合成するために，VocaWatcher とは完全に独立に使用することができる．一方，VocaWatcher は，合成した歌声に時刻同期した自然な顔動作を生成するために必ず VocaListener あるいは VocaListener2 とともに使用する必要がある．これら三つのシステムの位置づけを，第 1 図にまとめる．

2. VocaListener: 音高と音量を真似る歌声合成システム

VocaListener[4,5] は，お手本となる人間の歌声を用いて，その音高と音量を真似るようにインタラクティブに歌声合成できるシステムである．歌声の音響信号とその歌詞のテキストを与えると，楽譜情報を自動的に推定するため，音符を入力する必要がない利点ももつ．ヤマハの VOCALOID[3,12] に基づくさまざまな声質の歌声合成ソフトウェアが市販されているが，それらの歌声合

* 産業技術総合研究所
Key Words: singing information processing, singing synthesis, humanoid robot.



第 2 図 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

成パラメータ（音高と音量の表情付け用パラメータ）を推定できるため，同一の音高と音量，歌詞を持ち，声質だけが違うさまざまな歌声を容易に合成することができる．VocaListener による歌声は，お手本の歌声の自然な音高と音量が模倣されているため，時間のかかるパラメータ調整を手作業でしなくても，人間らしい表情豊かな歌声となる．

従来，人間らしい歌声を作るために，歌声合成に関するさまざまな研究がなされてきた [13–16]．近年では，波形接続方式 [12,17,18] や HMM（隠れマルコフモデル）合成方式 [19–21] といったコーパスベースの合成手法により，実用性の高い歌声合成システムが提案されている．これらはすべて「歌詞」と「楽譜情報（音符系列）」を入力として歌声を合成する方式であり，歌詞を入力として歌声を出力することから，このアプローチは lyrics-to-singing (text-to-singing) とよばれる [22]．自然性を増したり表情付けをしたりするために，音高や音量などの歌声合成パラメータをユーザが指定できるシステム [12] もあるが，その調整は容易でなく時間がかかっていた．一方，speech-to-singing[22] とよばれる別のアプローチでは，合成対象の歌詞を朗読した話声を入力とし，それを変換した歌声を出力するシステムが提案されている．もとの声質を保ったまま歌声が合成できるが，さまざまな声質を扱うことはできなかった．

それに対して VocaListener のアプローチは，歌声を入力として歌声を出力することから，われわれはこれを singing-to-singing と名付けた [4,5]．これに類似した先行研究として，Janer ら [23] は入力された歌声から音高と音量，ビブラートを推定し，それを正規化してそのまま歌声合成パラメータへ変換していた．しかし，同じパラメータを指定しても，歌声合成システムやその音源データ（声質の種類）といった歌声合成の条件が変わると，合成結果の音高と音量は指定通りにならない．これはたとえば，音源データごとに異なる音源波形（歌声の断片）が切り貼りされて合成されるため，その波形次第で変動するからである．そのため，歌声を真似る性能には限界があった．

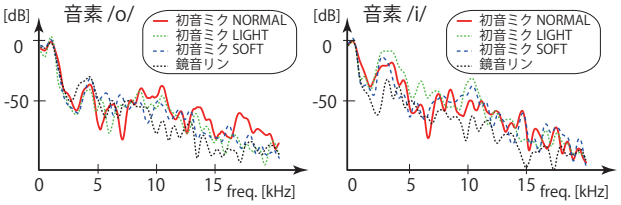
そこで VocaListener では，合成された歌声の音高と音量が入力と近くなるように，合成パラメータを反復更新することで，上記の条件の変化へ対処した．つまり，通常の歌声合成のように合成しっぱなしではなく，人があたかも何度も発声練習するかのように，合成音を再度取り込んで分析し，意図した通りでない部分のパラメータを補正して再度合成する処理を何度も反復する．これにより，歌い方を高精度に真似る歌声合成を実現した．第 2 図のように，まず，お手本の入力歌声を分析して音高，音量を推定し，歌声合成システムのパラメータに変換して合成する．つぎに，合成結果を分析して入力歌声と比較し，それらが近くなるようにパラメータを更新して再び合成する．これを十分に近くなるまで繰り返すことで，最終的な合成結果が得られる．

VocaListener には，歌詞からメロディの音符系列を得るために，お手本の歌声とその歌詞を自動的に対応付ける歌詞同期（歌詞アラインメント）機能がある．たとえば日本語の場合，歌詞の読み（ひらがな）の各文字（音節）が，歌声のどの区間に対応するかを推定し，その区間の音高をもつ一つの音符とする．そのためには音節（たとえば「か」）を構成する音素（/k/や/a/）の単位で音響的な特徴をモデル化した音響モデル（隠れマルコフモデル (HMM)）を，歌声専用に学習・適応させて独自に用意した．これを用いた Viterbi アラインメントによって時刻同期が実現できるが，さらにその結果の有声区間にずれがあれば自動的に微調整する工夫をした．それでも部分的に同期がずれて誤ることが起きてしまうため，誤っている箇所をユーザが指摘するだけで容易に訂正できる「ダメ出し」インタフェースも実現した．

さらに，お手本の歌唱力不足を補ったり，自分好みの表現へ変更したりできる歌唱力補正インタフェースも用意した．音高のずれを小さくする「調子はずれ補正機能」や声域を変える「音高トランスポーズ機能」，ビブラートなどを調整する「歌唱スタイルの変更機能」といった一連の機能から，好みに応じて選ぶことができる．

3. VocaListener2: 音高と音量だけでなく声色変化も真似る歌声合成システム

VocaListener2[6,7] は，お手本の歌声の音高と音量だけでなく，声色（こわいろ）の時間変化も真似ることができる歌声合成システムである．これは上記の音高と音量のみを真似する VocaListener の拡張であり，歌声とその歌詞のみを入力とする singing-to-singing のアプローチに基づいている．VocaListener2 を実現するために，VocaListener によって同一の音高と音量，歌詞をもち，声質だけが違うさまざまな歌声を合成し，それらの歌声を用いて，対象とする楽曲に固有の「声色空間」を構成する．この空間内で，お手本の声色の時間変化を軌



第 3 図　音素の違いと歌唱者の違いによるスペクトル包絡形状の違いの具体例

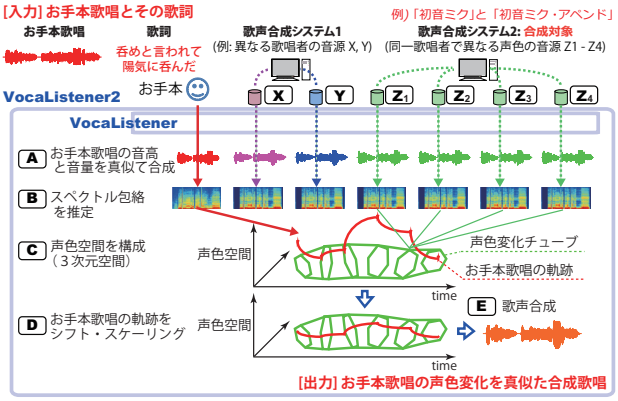
跡として表現し，その軌跡を真似るように合成結果を生成する．VocaListener では既存の歌声合成ソフトウェアのパラメータを制御していたが，VocaListener2 はそれでは実現できないため，スペクトル包絡を直接制御して合成する点が大きく異なる．

従来，異なる話者間の声質変換 [24,25] や感情音声合成 [26–28]，ラップ歌唱の声質変換 [29]，歌声合成結果の声質変換 [30]，二人の歌唱者による同一歌詞の歌声からの声質のモーフィング [31] などの声質操作に関連した研究はあったが，人間が歌唱中に意図的に変更する声色の変化を反映することはできなかった．一方，市販の歌声合成ソフトウェア VOCALOID[12] でも，歌声合成パラメータを各時刻で調整することで，合成結果の音響的特徴（スペクトルなど）をある程度操作できる．しかし，曲に合わせてこれらのパラメータを操作することは難しく，多くのユーザはこれらを変更しないか，変更するにしても曲ごとに一括で変更したり，大まかに変更したりしていた．

また，一部の市販の歌声合成ソフトウェアでは，同一歌唱者の声質で複数の歌唱スタイル（声色）の歌声を合成できるが，その中間の歌声は合成できなかった．たとえば，「初音ミク・アペンド (MIKU Append)」では，「初音ミク」(NORMAL) と同一歌唱者の個人性を保ったまま，DARK, LIGHT, SOFT, SOLID, SWEET, VIVID の 6 種類の声色で歌声合成できる．しかし，これらの声色をフレーズごとに切り替えながら合成することはできても，たとえば「LIGHT と SOLID の中間の声」は出せなかった．

そこで VocaListener2 では，声色の時間変化が反映されるように，歌声のスペクトル包絡形状を信号処理で直接制御して合成する．われわれは声色の違いを，(たとえば「初音ミク」と「初音ミク・アペンド」の) スペクトル包絡形状の違いとして扱う．第 3 図に示すように，スペクトル包絡形状の違いには，声色の違い（NORMAL と LIGHT）だけでなく，音素の違い（/o/ と /i/）や歌唱者の違い（初音ミクと鏡音リン）も含まれる．しかし，VocaListener によって音高と音量，音素が一致した状態のスペクトル包絡形状を比較すれば，歌唱表現としての声色変化だけを調べることができる．

処理の流れを第 4 図に示す．全体は，VocaListener，歌声分析（A〜D），歌声合成（E）で構成される．

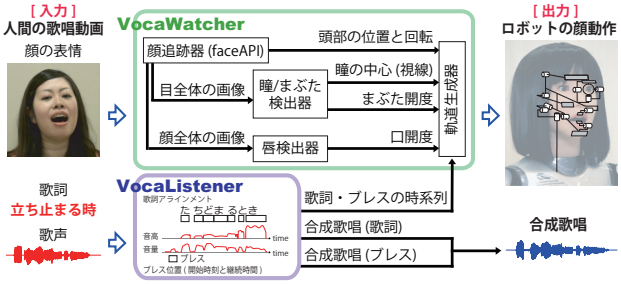


第 4 図　VocaListener2 の処理概要．人間の歌声の音高・音量・声色変化を真似るように歌声合成する

まず，お手本の歌声の音高と音量を真似た歌声を，VocaListener を用いて複数生成する（A）．これによって，お手本歌唱と時刻が同期した多様な声質の歌声が得られる．続いて，それぞれの歌声を分析し，音高による影響を除去したスペクトル包絡を推定する（B）．そのようにして得た同一の音高，音量，音素におけるスペクトル包絡の違いは声色の違いを表すので，部分空間法を用いて，より低次元（現在の実装では 3 次元）で声色を的確に表現する**声色空間**を構成する（C）．これにより，各時刻において，ある歌声の声色はこの空間内の一点で表すことができるので，お手本の歌声や合成された各歌声は，この空間内で時間変化する軌跡として表現できる．声色空間を作る際には可能な限り多くの音源データ（声質の種類）による合成結果を用いる．一方，合成する際には，同一歌唱者の声色が異なる一連の合成結果のみ（Z_1–Z_4）を選んで，それら複数の軌跡を含むような時々刻々と変化する多面体（ポリトープ）が声色変化可能な領域であると考える．その多面体の時間軌跡を**声色変化チューブ**とよぶ．したがって，声色空間の別の場所に存在するお手本歌唱の軌跡を，声色変化チューブ内になるべく入るようにシフト・スケーリングして調整し（D），その調整後の位置に対応するスペクトル包絡を各時刻において生成することで，歌声合成結果を得る（E）．

4.　VocaWatcher：歌唱時の表情を真似るロボット顔動作生成システム

VocaWatcher[8,9] は，単一の家庭用ビデオカメラで上半身が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するシステムである．顔にマーカーを一切付与せずに，映像中の人間の口，目，首の動きを推定し，それを真似るようにヒューマノイドロボット「HRP-4C 未夢」[32] の動作を制御する．HRP-4C は，人間の女性に近い外観（身長 160cm，体重 46kg，44 自由度）と動作性能をもち，表情の制御が可能という特長をもつ．



第 5 図 VocaWatcher の処理概要．人間の歌い手を真似るようにロボットの自然な顔動作を生成する

VocaListener あるいは VocaListener2 によって得られた合成歌唱の歌詞の時系列を用いることで，歌詞と正確に同期したロボットの顔動作が生成できる．

音楽はヒューマノイドロボットにとって重要で魅力的な応用例であるため，楽器を演奏するさまざまなロボット [33–35] が開発されてきた．歌声合成技術を用いて歌うヒューマノイドロボット [36–38] も開発されたが，動作を手作業で振り付けて制御されていた．われわれの HRP-4C も，歌声合成ソフトウェア VOCALOID に，キーポーズと状態遷移モデルなどに基づく動作生成を組み合わせて歌唱するデモンストレーションを，技術展示会 CEATEC JAPAN 2009 において展示した [38]．しかし，いずれの場合も，自然性には限界があった．一方，コンピュータグラフィックスですでに実証されているように，人間の動きを真似るモーションキャプチャは，ロボットの自然な動きを作るうえでも有効である．しかし，顔へのマーカー付与 [39] や，特定個人用の顔モデルの学習 [40] が必要という問題があった．しかも，顔動作の制御に音響信号処理は使用されていなかった．

そこで VocaWatcher では，第 5 図 のように画像処理と音響信号処理を組み合わせることで，ヒューマノイドロボット HRP-4C のより自然な歌唱を，歌声とそれに同期した顔の動きの両方の観点で実現する．自然な歌声に関しては，前述の VocaListener あるいは VocaListener2 により，人間の歌唱を収録した動画中の歌声を模倣して合成する．ただし，VocaListener2 は実際にはまだロボットと組み合わせて使用していない．それと同時に，その動画中の顔画像を分析して，歌声に同期したロボットの口，目，首の動きを生成する．VocaWatcher では，顔のマーカーや複数のカメラが不要で，かつ，VocaListener が推定した各音節の時刻情報を活用して高精度な制御ができる利点をもつ．

処理の流れを第 5 図に示す．VocaWatcher はまず，市販の顔検出ソフトウェア faceAPI (Seeing Machine 社) を用いて，頭部の位置と回転（姿勢）を推定する．つぎに，画素よりも細かいサブピクセルを考慮した独自の画像処理手法によって瞳の中心（視線）とまぶた開度（まばたき）を検出し，パーティクルフィルタに基づく時間変化追跡手法によって口開度（上唇と下唇間の距離）を

(a) 歌い出しで /ta/ を発声

(b) 目を閉じた状況で /ru/ を発声

(c) 長い /ra/ を発声

第 6 図 お手本となる人間の歌い手の顔（左）とそれを真似るように顔動作を制御した HRP-4C の顔（右）の具体例

検出する．そして，それらの検出結果に基づいて，首（3 関節），目（2 関節），口（4 関節）の動きを生成する．さらに口に関しては，母音と撥音，ブレスに対応する関節角度（キーポーズ）をあらかじめ定めておき，歌声中のそれぞれの開始タイミングに，口がその形状へなめらかに変化するように制御する．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 をブレス音も真似て合成するように拡張した．

第 6 図に，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．このデモンストレーションは，技術展示会 CEATEC JAPAN 2010 の展示で初めて公開したが [9]，その際にはより魅力的な歌唱となるように，音楽に併せて腕も動くように手作業で振り付けた．

5. おわりに

本解説では，メインボーカルが人間でない音楽制作のために，容易に表情豊かな歌声が合成できるシステムを実

現し，ヒューマノイドロボットの音楽分野への応用を促進するために，自然な歌唱の顔動作が生成できるシステムを実現するわれわれの研究を紹介した．これらのデモンストレーションの動画は，http://staff.aist.go.jp/t.nakano/システム名/ の「システム名」を，VocaListener，VocaListener2，VocaWatcher のいずれかで置き換えれば閲覧できる．

楽音合成（シンセサイザー）が登場して普及し，ポピュラー音楽制作で欠かせない存在となったのと同様，歌声合成がいつの日か普及することは歴史的必然である．楽音合成も当初は自然な楽器音と容易に区別がつき，それがゆえに独自の表現も生まれたが，今日では非専門家には区別のつかない高いクオリティとなり，ポピュラー音楽の大半で用いられている．歌声合成がそうならない理由はない．不確定なのは，それが今回のタイミングで起きるのか，より技術が進歩した未来に起きるのか，の時期の違いだけである．

そして，その歌声合成が広く普及する未来へ向けて，さまざまなユーザインタフェースの選択肢を用意しておくことが重要だと，われわれは考えている．前述したように現在は lyrics-to-singing (text-to-singing) のアプローチが主流だが，楽譜やピアノロールで編集をするのが難しいユーザがいる．そこで，音楽家にとって，手で弾いた演奏をリアルタイムに MIDI で入力するインタフェースが直感的で不可欠となったように，楽曲を歌えるユーザや伴奏なしの独唱に基づいて合成したいユーザにとっては，われわれの singing-to-singing のアプローチが，容易で効率的なインタフェースとして今後不可欠になる可能性がある．同様に，ロボットによる歌唱を個人が自分好みに制御して楽しむ未来では，VocaWatcher のような顔の動きで容易に入力できるインタフェースの方が，手作業で個々の動作を指定するよりも普及する可能性がある．そうした時間のかかる手作業での調整が不要で直感的な入力手段は，ユーザが楽曲のメッセージや意図を伝えることに集中するうえで重要である．

本研究では，お手本の「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．もちろん，単に模倣するだけでは自由な表現を創り出すうえで限界がある．しかし，ひとたび，歌声合成やロボット動作生成のための制御パラメータ空間内で極めて自然な表現が得られれば，将来の次の段階として，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が可能になる．これにより，模倣を越えた新たな表現へつなげていければと考えている．

VocaListener や VocaListener2 によって合成された歌声や，VocaWatcher によって生成された顔動作は，人間らしく自然なため，少し視聴しただけでは人間と区別がつかないぐらいの良い印象を受けた人がいた．その一方で，顔の動作や声の質，皮膚や顔形状などの見た目に関して，まだ不自然さが残るため，気味の悪さを感じる人もいた．これは，「不気味の谷」と名付けられた有名な現象で，人間から遠くデフォルメされていると親近感を得やすくても，人間に近づく過程で違和感を感じ，親近感を持ちにくくなる感情を説明した言葉である．しかし，自然な表現に関する技術が漸次的に進歩すれば，どこかの段階で「不気味の谷」にさしかかるのは不可避である．重要なのは，まずそこにたどり着き，そのうえで技術の力で乗り越えることであるとわれわれは信じている．

謝　辞

本研究の一部は，JST CREST による支援を受けた．お手本歌唱として RWC 研究用音楽データベース [41] を使用し，ブレス検出のモデル学習のために AIST ハミングデータベース [42] を使用した．本研究の合成結果や研究活動全般について，ニコニコ動画などのさまざまな場でコメントし，議論して下さった方々に感謝する．

(2012 年 2 月 1 日受付)

参 考 文 献

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

[4] 中野，後藤：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；情処学論，Vol. 52, No. 12, pp. 3853–3867 (2011)

[5] T. Nakano and M. Goto: VocaListener: A singing-to-singing synthesis system based on iterative parameter estimation; *Proc. of SMC 2009*, pp. 343–348 (2009)

[6] 中野，後藤：VocaListener2: ユーザ歌唱の音高と音量だけでなく声色変化も真似る歌声合成システムの提案；情処研報音楽情報科学，Vol. 2010-MUS-86, No. 3, pp. 1–10 (2010)

[7] T. Nakano and M. Goto: VocaListener2: A singing synthesis system able to mimic a user's singing in terms of voice timbre changes as well as pitch and dynamics; *Proc. of ICASSP 2011*, pp. 453–456 (2011)

[8] 中野，後藤，梶田，松坂，中岡，横井：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；情処研報音楽情報科学，Vol. 2012-MUS-94, No. 6, pp. 1–10 (2012)

[9] S. Kajita, T. Nakano, M. Goto, Y. Matsusaka, S. Nakaoka and K. Yokoi: VocaWatcher: Natural singing motion generator for a humanoid robot; *Proc. of IROS 2011* (2011)

[10] 後藤，齋藤，中野，藤原：解説 “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)
[11] M. Goto, T. Saitou, T. Nakano and H. Fujihara: Singing information processing based on singing voice modeling; *Proc. of ICASSP 2010* (2010)
[12] H. Kenmochi and H. Ohshita: VOCALOID – Commercial singing synthesizer based on sample concatenation; *Proc. of Interspeech 2007*, pp. 4011–4010 (2007)
[13] P. Depalle, G. Garcia and X. Rodet: A virtual castrato; *Proc. of ICMC'94*, pp. 357–360 (1994)
[14] P. R. Cook: Identification of Control Parameters in an Articulatory Vocal Tract Model, with Applications to the Synthesis of Singing, PhD Thesis, Stanford University (1991)
[15] P. R. Cook: Singing voice synthesis: history, current work, and future directions; *Computer Music Journal*, Vol. 20, No. 3, pp. 38–46 (1996)
[16] J. Sundberg: The KTH synthesis of singing; *Advances in Cognitive Psychology, Special issue on Music Performance*, Vol. 2, pp. 131–143 (2006)
[17] 吉田，中嶌：歌声合成システム: CyberSingers；情処研報 音声言語情報処理 99-SLP-25-8，pp. 35–40 (1998)
[18] J. Bonada and X. Serra: Synthesis of the singing voice by performance sampling and spectral models; *IEEE Signal Processing Magazine*, Vol. 24, No. 2, pp. 67–79 (2007)
[19] 酒向，宮島，徳田，北村：隠れマルコフモデルに基づいた歌声合成システム；情報処理学会論文誌，Vol. 45, No. 7, pp. 719–727 (2004)
[20] K. Saino, H. Zen, Y. Nankaku, A. Lee and K. Tokuda: An HMM-based singing voice synthesis system; *Proc. of Interspeech 2006*, pp. 1141–1144 (2006)
[21] K. Saino, M. Tachibana and H. Kenmochi: A singing style modeling system for singing voice synthesizers; *Proc. of Interspeech 2010*, pp. 2894–2897 (2010)
[22] T. Saitou, M. Goto, M. Unoki and M. Akagi: Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices; *Proc. of WASPAA 2007*, pp. 215–218 (2007)
[23] J. Janer, J. Bonada and M. Blaauw: Performance-driven control for sample-based singing voice synthesis; *Proc. of DAFx-06*, pp. 41–44 (2006)
[24] T. Toda, A. Black and K. Tokuda: Voice conversion based on maximum likelihood estimation of spectral parameter trajectory; *IEEE Trans. on ASLP*, Vol. 15, No. 8, pp. 2222–2235 (2007)
[25] F. Villavicencio and E. Maestre: GMM-PCA based speaker-timbre conversion on full-quality speech; *Proc. of SSW-7*, pp. 56–61 (2010)
[26] O. Türk and M Schröder: A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis; *Proc. of Interspeech 2008*, pp. 2282–2285 (2008)
[27] T. Nose, M. Tachibana and T. Kobayashi: HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation; *IEICE Trans. on Information and Systems*, Vol. E92-D, No. 3, pp. 489–497 (2009)
[28] Z. Inanoglu and S. Young: Data-driven emotion conversion in spoken English; *Speech Communication*, Vol. 51, No. 3, pp. 268–283 (2009)
[29] O. Türk, O. Büyük, A. Haznedaroglu and L. M. Arslan: Application of voice conversion for cross-language rap singing transformation; *Proc. of ICASSP 2009*, pp. 3597–3600 (2009)
[30] F. Villavicencio and J. Bonada: Applying voice conversion to concatenative singing-voice synthesis; *Proc. of Interspeech 2010*, pp. 2162–2165 (2010)
[31] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi and H. Banno: Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown; *Proc. of ICASSP 2009*, pp. 3905–3908 (2009)
[32] K. Kaneko, F. Kanehiro, M. Morisawa, K. Miura, S. Nakaoka and S. Kajita: Cybernetic human HRP-4C; *Proc. of Humanoids 2009*, pp. 7–14 (2009)
[33] I. Kato, S. Ohteru, K. Shirai, T. Matsushima, S. Narita, S. Sugano, T. Kobayashi and E. Fujisawa: The robot musician WABOT-2 (Waseda Robot-2); *Robotics*, Vol. 3, pp. 143–155 (1987)
[34] K. Chida, I. Okuma, S. Isoda, Y. Saisu, K. Wakamatsu, K. Nishikawa, J. Solis, H. Takanobu and A. Takanishi: Development of a new anthropomorphic flutist robot WF-4; *Proc. of ICRA 2004*, pp. 152–157 (2004)
[35] T. Mizumoto, H. Tsujino, T. Takahashi, T. Ogata and H. Okuno: 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)
[36] Y. Kuroki, M. Fujita, T. Ishida, K. Nagasaka and J. Yamaguchi: A small biped entertainment robot exploring attractive applications; *Proc. of ICRA 2003*, pp. 471–476 (2003)
[37] K. Murata, K. Nakadai, K. Yoshii, R. Takeda, T. Torii, H. G. Okuno, Y. Hasegawa and H. Tsujino: A robot singer with music recognition based on real-time beat tracking; *Proc. of ISMIR 2008*, pp. 199–204 (2008)
[38] M. Tachibana, S. Nakaoka and H. Kenmochi: A singing robot realized by a collaboration of VOCALOID and cybernetic human HRP-4C; *Proc. of InterSinging 2010*, pp. 9–14 (2010)
[39] F. Wilbers, C. Ishi and H. Ishiguro: A blendshape model for mapping facial motions to an android; *Proc. of IROS 2007*, pp. 542–547 (2007)
[40] P. Jaeckel, N. Campbell and C. Melhuish: Facial behavior mapping — From video footage to a robot head; *Robotics and Autonomous Systems*, Vol. 56, pp.

1042–1049 (2008)
[41] 後藤，橋口，西村，岡：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；情処学論，Vol. 45, No. 3, pp. 728–738 (2004)
[42] 後藤, 西村:AIST ハミングデータベース: 歌声研究用音楽データベース；情処研報 音楽情報科学 2005-MUS-61-2, pp. 7–12 (2005)

著 者 略 歴

後藤 真孝

1998 年早稲田大学大学院理工学研究科博士後期課程修了．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 上席研究員 兼 メディアインタラクション研究グループ長．統計数理研究所客員教授，筑波大学大学院 准教授（連携大学院），IPA 未踏 IT 人材発掘・育成事業プロジェクトマネージャーを兼任．博士（工学）．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 長尾真記念特別賞など 27 件受賞．

中野 倫靖

2003 年図書館情報大学卒業，2008 年筑波大学図書館情報メディア研究科博士後期課程修了．現在，産業技術総合研究所研究員．博士（情報学）．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．情報処理学会，日本音響学会各会員．

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 足歩行ロボットなど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 より組織改変に伴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）．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受賞．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）．

松坂 要佐

2003 年早稲田大学大学院理工学研究科修了．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，現在に至る．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．博士（工学）．

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．2006 年 4 月より産業技術総合研究所知能システム研究部門研究員．2011 年 11 月より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の研究に従事．博士 (情報理工学)．2008 年度日本ロボット学会論文賞，IROS2010 Best Paper Finalist などを受賞．日本ロボット学会会員．

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 年〜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）兼任．人間型ロボットなどの研究に従事．博士（工学）．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，IEEE などの会員．共著書に，ヒューマノイドロボット（オーム社）ほか．