VocaRefiner: 歌を歌って歌い直して統合できる新しい歌声生成インタフェース

中野 倫靖　　後藤 真孝*

概要.　本稿では，音楽制作における歌声パートの生成において，歌手が一度の歌唱のみでは望んだ歌い方を得られない状況を想定し，何度も歌ったり気に入らない箇所だけを歌い直すことで，それらを統合して一つの歌声を生成できるインタラクティブシステム VocaRefiner を提案する．従来，部分的に歌い直して置換したり，歌声の音高や音量を補正したり，声質（音色）の変換やモーフィングをしたりすることはできたが，同一人物が断片的に複数回歌唱して，それらを統合する歌声生成のインタラクションは考えられていなかった．VocaRefiner は，複数の歌声から優れた一つに代表させて置換するのではなく，それぞれを音の三要素である音高・音量・音色に分解し，その要素単位で置換する．さらに，それらを時間軸上で伸縮したり，要素を個別に補正したりするインタラクションも可能とする．そうすることで，一度歌った歌に音高だけをハミング等の歌詞なし歌唱で入力し直したり，うまく歌えない箇所はマウスで音高に関する情報を入力して歌声生成したり，本来は速い歌唱をゆっくり歌えたりできるようになる．

1　はじめに

音楽制作におけるより手軽な歌声生成を目指して，現在の歌声生成の限界を超えるためのインタフェースを提案する．歌声は音楽の重要な要素であり，音楽は産業・文化の両面で主要なコンテンツの一つである．特にポピュラー音楽では歌声を中心に音楽を聴く人が多く，歌声の生成を極めることは，音楽制作において有用である．さらに，歌声は音の三要素である音高・音量・音色の全てが複雑に変化する時系列信号であり，特に音色は歌詞の音韻が次々と変化するため他の楽器音の生成よりも技術的に難易度が高い．したがって，このような歌声を効率的に生成できる技術やインタフェースの実現は学術的にも意義がある．

現在，歌声を生成するためには，まず「人間が歌う」か「歌声合成技術（歌声合成用パラメータの調整）によって人工的に生成する」ことで，基となる歌声の時系列信号を得る必要がある．本稿ではこのように，歌声合成技術による生成だけでなく，人間の歌唱も一種の生成と見なして，包括して「歌声生成」と呼ぶ．さらに，必要に応じてそれらを切り貼りしたり，信号処理技術等によって時間伸縮や変換をしたりしながら「編集する」ことで，最終的な歌声を得る場合もある．したがって，歌唱力がある人，歌声合成のパラメータ調整が得意な人，歌声を上手に編集できる技術を持っている人は「歌（声生成）が上手い」と言える．このように歌声生成は，高い歌唱力や高度な専門知識，手間のかかる作業が必要とされ，前述のようなスキルがない人々にとっては，質の高い歌声を自在に生成することはできなかった．

このような，歌声を生成するためのスキルの種類を増やすことができれば，これまで歌をうまく生成していた人の可能性を拡げるだけでなく，これまで歌をうまく生成できなかった人も，音楽制作に加わることができる可能性がある．すなわち，質の高い様々な音楽コンテンツの増加につながると考えられる．

本稿では，現在の歌声生成の主流である「人間による歌声生成」と「計算機による歌声生成（合成及び変換）」の限界を，インタラクションによって超えるインタフェース VocaRefiner を提案する．VocaRefiner は，歌手が何度も歌ったり気に入らない箇所だけを歌い直し，それらを統合して一つの歌声を生成するインタラクティブシステムである．これによって歌声を効率的に生成し，インタラクションに基づく新しい「歌声生成スキル」へつなげる．

2　歌声生成における限界の超え方

本章ではまず，歌声生成における人間と計算機による歌声生成それぞれの利点と限界を述べる．次に，制作対象の曲を思い通りの歌い方で歌っている人間の歌声を活用することで，両者の利点を生かして限界を超える方法について考察する．

2.1　人間の歌声生成能力とその限界

多くの人は歌唱力を問わなければ容易に歌うことができ，その歌声は人間らしくて自然性が高い．また，既存の歌を自己流に歌い回しを変える表現力を持っている．特に，歌唱力がある人であれば，音楽

* Tomoyasu Nakano and Masataka Goto, 産業技術総合研究所（AIST）

的に質の高い歌声を生成することが可能で，聴く人に感動を与えることができる．

しかし，過去に歌った歌を再現してもう一度歌ったり，自身の限界よりも声域が広い歌を歌ったり，歌詞が速い歌を歌ったり，自分の歌唱力を超えた歌を歌ったりすることには困難を伴う場合がある．

2.2 計算機の歌声生成能力とその限界

計算機による歌声生成の利点は，多様な声質の合成が行えて，一度合成した歌唱の表現を再現できる点にある．また人間の歌声を，音の三要素である音高・音量・声質に分解して，それぞれを個別に制御して変換できる．特に歌声合成ソフトウェアを使う場合，ユーザは歌唱しなくても歌声を生成できるために，場所を選ばずにどこでも生成できて，さらに何度も聴取しながら表現を少しずつ変更できる．

しかし，人間の歌声と区別がつかないような自然な歌声を自動的に生成したり，想像力によって新たな歌声表現を生み出したりすることは，一般的には困難である．例えば，自然な歌声で合成するためには手作業での精密なパラメータ調整が必要で，多様で自然な歌唱表現を得るのは容易でない．また，合成と変換のいずれも，元となる歌声（歌声合成データベースの音源や声質変換前の歌声）の品質によっては，合成・変換後に良い品質が得られにくい．

2.3 限界の超え方

本研究では複数回歌った歌唱をインタラクティブに統合することで，前述のような限界を超えることを考える．そのために，人間の歌声生成と，計算機による歌声生成両者の利点を付与する．具体的には人間の歌声を計算機で処理（変換）する方法が，第一に考えられる．デジタル録音によって劣化少なく再現でき，信号処理技術によって肉体的な制約を超えた変換も行える．第二に，計算機での歌声合成を人間の歌声によって制御することが考えられる．例えば，ユーザ歌唱をからその歌い方を真似た歌声合成パラメータを推定する VocaListener [21] が提案されており，より人間らしい歌声の合成につながった．

しかし，どちらの場合でも，前述した信号処理技術の限界（合成と変換の品質が基の歌声に依存する）によって，より質の高い歌を生成するためには，ミスや乱れのない歌声が得られることが望ましい．そのためにはほとんどの場合，たとえ歌唱力が高くても納得のいくまで歌い直す必要があるため，何度も歌い直して録音した後，それを切り貼りして優れた部分のみを統合する処理が必要となる．しかし従来，そういった複数回歌われた歌声の扱いを視野に入れた研究はなかった．

そこで本稿では，人間と計算機の歌声生成を融合させるアプローチに基づき，人間が複数回歌った歌唱を扱うためのインタラクション機能を持つ歌声生成インタフェース VocaRefiner を提案する．

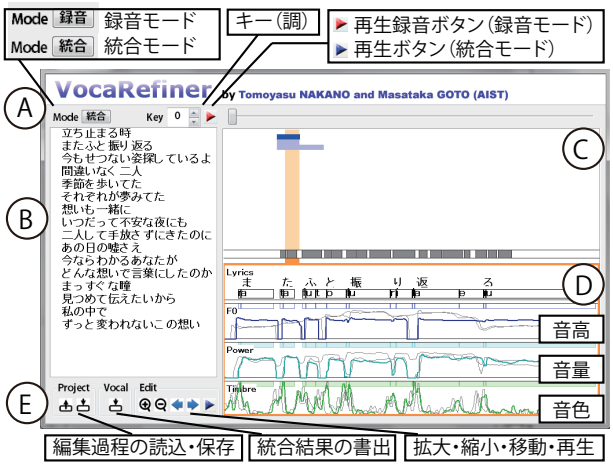

図 1. VocaRefiner の実行画面（統合モード）

3 VocaRefiner: 歌声の統合によって限界を超える歌声生成インタフェース

VocaRefiner の起動画面を**図 1** に示す．VocaRefiner には，歌唱の伴奏となる背景音楽に時刻同期してユーザの歌唱を録音する「録音モード」と，録音モードで録音した複数の歌唱を統合するための「統合モード」の二種類を実装した．ユーザは，まず，歌詞のテキストファイルと背景音楽の音響信号ファイルを入力してから，それらに基づいて歌唱して録音する．ここで，既に背景音楽が用意されていると仮定する[1]．また，歌詞のテキストファイルには，漢字かな交じりの歌詞と，背景音楽中における歌詞の各文字の時刻，及び読み仮名が含まれているものとする．録音後，歌声を確認・編集しながら統合する．

3.1 インタフェースの概要

録音モードと統合モードとの相互変更は，画面左上Ⓐのモード変更ボタンで行う．

まず，録音モードでは「歌詞ウィンドウⒷ」に，入力された時刻情報付の歌詞から，歌詞のテキストが表示される．背景音楽の再生は，画面中央上部（Ⓐの右）にある「再生録音ボタン（録音モード）」もしくは「再生ボタン（統合モード）」によって行う．歌声は楽曲の再生と同時に常に録音されており，「録音統合ウィンドウⒸ」にその録音区間を示す矩形（青）が画面右上の再生バーと同期して表示される．再生録音の時刻は，再生バーや歌詞中の任意の文字のダブルクリックでも指定できる．さらに，背景音楽の

[1] 背景音楽にはボーカルやガイドメロディー音が含まれている方が歌いやすい．ただし，歌いやすいようにミックスバランスは通常と違っていてもよい．

図 2. 録音モードのインタラクション

音高を周波数軸上にシフトさせることで，キー（音楽の調）を変更できるキー変更ボタンがある．

統合モードでは，分析結果Ⓓの表示範囲を拡大･縮小したり，左右に動かしながら編集･統合Ⓔを行う．

3.2 録音モードのインタラクション

実際に歌唱を録音する状況を考えた場合，歌を短時間で可能な限り多く録音して，後でそれらを吟味した方が効率的な場合がある．例えば，スタジオを借りていて時間制限がある場合等である．そこで録音モードでは，歌唱することに集中して効率的に録音するために，楽曲の再生と同時に常に録音状態にし，必要最低限なインタラクションのみを行う．

以下に，録音モードのインタラクションをユーザのアクションと VocaRefiner のリアクションに分けて述べて，その様子を**図 2** に示す．

3.2.1 ユーザによるアクション（録音）

ユーザによるアクションは，基本的には「再生・録音時刻の指定」と「キーの変更」であるが，歌声を客観的に聴くために「録音歌唱の再生」もできる．歌唱は歌詞に沿った「音素付き」で歌うことを前提として処理を行うが，例えば，ハミングや楽器音で音高入力をした場合には，統合モードで修正する．

再生録音機能 「再生」をイメージさせる三角形状のボタンを「録音」をイメージさせる赤い色でデザインした（図 1）．本研究ではこれを「再生録音ボタン」と呼び，背景音楽の再生とそれに同期した録音を行う．また，再生バーで再生時刻を決めて録音することもできる．

歌詞による頭出し再生録音機能 歌詞中の文字のダブルクリックによって，その文字が始まる時刻の頭出しを行う．従来，時刻情報付きの歌詞を再生中にカラオケ表示のようにして楽しむ目的で利用することはあったが，歌声の録音に用いられた例はなかった．本研究では，歌詞は音楽中の時刻を指定できる一覧性の高い有用な情報として利用する．

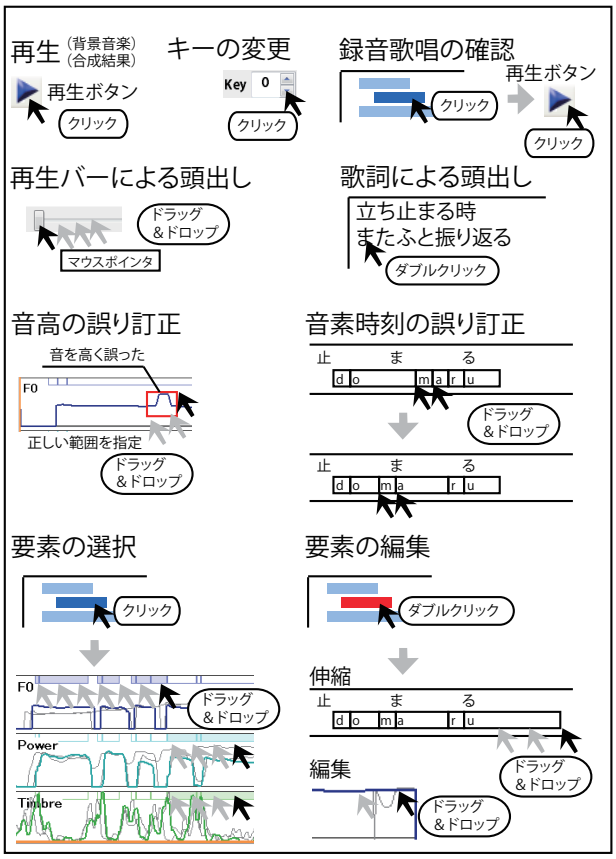

図 3. 統合モードのインタラクション

特定の歌詞を自由に歌う機能 実際の歌詞の時刻情報を無視して，本来は速い歌唱をゆっくり歌ったり，そのままでは歌うのが難しい場合に自分なりに歌ったりできる．歌詞をマウスドラッグで選択した後，再生録音ボタンを押すことで，選択された歌詞の時間範囲を歌っていると仮定して録音する．

キーを変更して再生録音する機能 キーを変更して伴奏を聴取でき，それに合わせて歌唱できる．

録音した歌唱の確認機能 聴きたい歌唱をクリックして選択し，画面下の再生ボタンで再生する．

3.2.2 VocaRefiner のリアクション（録音）

VocaRefiner は歌詞の読み仮名を用いて，歌詞と歌声の自動的な対応付けを行う．対応付けでは，再生された時刻付近の歌詞が歌われていると仮定し，特定の歌詞で自由に歌う機能を用いた場合は，選択された歌詞を仮定する．また，歌声を音高・音量・声質の三要素に分解する．

具体的には，一つの録音が終わる毎に，バックグラウンド処理によって音高・音量を推定する．ここで，統合モードで必要となる声質に関する全情報の推定には時間を要するため，歌詞の時刻を推定するために必要な情報のみを計算する．全ての録音が終わり，統合モードで情報が必要になる時点で，声質情報の推定を開始する旨をユーザに提示する．

3.3 統合モードのインタラクション

録音時に多数歌われた箇所は，歌唱に納得がいかずに歌い直した可能性がある．そこで統合モードにおける初期状態では，後に録音された歌声が選択されている．ただし，全ての音が録音されているために，単純に一番最後の録音を選択しただけでは，無音で上書きさせる可能性がある．そこで，自動的に対応付けられている音素の時刻情報に基づいて，歌唱部分のみから録音の順番を判断する．しかし，自動対応付けで 100% の精度を得ることは現実的ではないため，誤りがあった場合にはユーザが修正する．

そのような状況を踏まえ，統合モードのインタラクションをユーザのアクションと VocaRefiner のリアクションに分けて述べて，その様子を**図 3** に示す．

3.3.1 ユーザによるアクション（統合）

ユーザによるアクションは,「自動推定結果の誤り訂正」,「統合（要素の選択と編集）」であり，録音とその分析結果，変換した歌声を視聴しながら行う．まず，音高と音素時刻の推定には，誤りが発生する可能性があるため，その場合にはここで訂正する．また再度，録音モードに戻って歌声を追加することも可能である．誤りを訂正した後，音素単位で歌声要素を選択したり編集したりして統合する．

誤り訂正：音高推定結果 音高の誤りには，マウスのドラッグ操作で音高の範囲を時間・音高（周波数）で指定して再推定する [21].

誤り訂正：音素時刻推定結果 音素の時刻の誤り訂正に関しては，録音モードでのインタラクションで既におおよその時刻と音素が与えられているために誤りが少ない．そこで現在の実装では，マウスによる微調整で誤りを訂正する．

また，推定結果の音素が足りない場合や多すぎる場合には，マウス操作で追加・削除を行う．

統合：要素の選択と編集 前述したように，初期状態では後に録音された要素が選択されているが，それ以前の要素を選択することもできる．

また，音素の長さを伸縮させたり，音高・音量をマウス操作で書き換えたりして編集できる．

3.3.2 VocaRefiner のリアクション（統合）

声質の推定が音高に依存しているため，訂正された誤り情報に基づき，音高・音量・声質を再推定する．また，統合された全時刻の三要素の情報から歌声の波形を合成する．

このように統合して得られた人間の歌声に基づいて，その歌い方を真似るように，特定の歌声合成データベースの声質で合成したい場合には，VocaListener [21] を使用するとよい．

4 VocaRefiner の実現方法

VocaRefiner を実現するための信号処理技術について，以下に述べる．録音・再生する音は，44.1kHz, 16bit のモノラル信号を用いた．

4.1 歌詞への時刻情報及び読み仮名の付与

漢字仮名交じりの歌詞のテキストファイルに対し，その読み仮名と時刻情報を付与しておく必要がある．手作業も可能だが，本研究では正確さと手軽さを考慮して，事前に歌詞のテキストと仮歌を用意し，VocaListener [21] を用いて，形態素解析と信号処理による歌詞のアラインメントを行った．

仮歌は，音素の発音時刻さえ正しければ良く，録音の品質が多少低くても，無伴奏歌唱であれば推定結果に影響は少ない．ここで，形態素解析の結果や，歌詞アラインメントに誤りがあった場合，VocaListener の GUI によって正しく訂正した．

4.2 背景音楽のキー（音楽の調）の変更

背景音楽のキーを変更するためには，フェーズボコーダ等 [9] で実現でき，リアルタイム実装も可能である．ただし本稿では，各キーに変更した音源を事前に作成し，その再生を切替えるように実装した．

4.3 録音された歌声と歌詞音素との自動対応付け

VocaListener [21] と同様の条件で対応付けを行った．具体的には，Viterbi アラインメントによって自動的に推定し，音節境界に短い無音（short pause）が入ることを許容した文法を用いた．また音響モデルには，連続音声認識コンソーシアムで頒布されている 2002 年度版の不特定話者 monophone HMM [19] を歌声に適応させて使用した[2]．音響モデル適応の際のパラメータ推定手法としては，MLLR（Maximum Likelihood Linear Regression）と MAP 推定（Maximum A Posteriori Probability）を組み合わせた MLLR-MAP [2] を用いた．特徴抽出と Viterbi アラインメントでは 16 kHz にリサンプリングした歌声を用い，MLLR-MAP による適応は HTK Speech Recognition Toolkit [8] で行った．

[2] 歌声のみで学習した HMM [21] も使用可能だが，話すように歌うことも考慮してこちらの HMM を用いた．

4.4 歌声の音の三要素への分解と合成

歌声の音高となる基本周波数（以下，F_0 と呼ぶ）の推定には，入力信号中で最も優勢な（パワーの大きい）高調波構造を求める手法 [16] で求めた値を初期値とした．16 kHz にリサンプリングした歌声を用い，1024 点のハニング窓で分析した．

さらに，その値に基づいて，元の歌声を F_0 適応させたガウス窓（分析長が $3/F_0$ の長さ）でフーリエ変換した後，その 10 倍音までの振幅スペクトルに，F_0 の整数倍の各倍音をそれぞれガウス分布の平均とする GMM (Gaussian Mixture Model) を EM (Expectation Maximization) アルゴリズムによってフィッティングさせて，F_0 推定の時間分解能と精度を向上させた．

また，音色（声質）の情報としてスペクトル包絡を推定するために，ソース・フィルタ分析を行った．本稿では，F_0 適応多重フレーム統合分析法 [22] によってスペクトル包絡と群遅延を推定し，分析と合成を行った．

5 関連する技術及び研究

従来の歌声生成に関しては，人間の歌声に加えて，近年では市販の歌声合成ソフトウェアが注目を集め，楽しむリスナーも増加している [15]．

歌声合成では，「歌詞」と「楽譜（音符系列）」を入力として歌声を合成する text-to-singing (lyrics-to-singing) 方式が主流であり，市販のソフトウェアでは，品質の高さから波形接続方式 [1, 4] が用いられているが，HMM（隠れマルコフモデル）合成方式 [12, 14] も利用され始めている．さらに歌詞のみを入力として自動作曲と歌声合成を同時に行うシステムも公開されており [3]，声質変換によって歌声合成を拡張する研究もある [7]．

一方，合成対象の歌詞を朗読した話声から，その声質を保ったまま歌声に変換する speech-to-singing 方式 [6, 11] やお手本の歌声を入力として，その音高や音量等の歌唱表現を真似るように歌声合成する singing-to-singing 方式 [21] が研究されている．

以上のようにして得られた歌声は，DAW (Digital Audio Workstation) 等を用いることで，切り貼りや信号処理を伴った時間軸伸縮や音高補正等が行える．その他，声質変換 [7,13,20] や音高と声質のモーフィング [5, 10]，高品質な実時間音高補正 [23] が研究されている．

また，楽器の MIDI シーケンスデータの生成において，リアルタイムの演奏入力が困難なユーザでも，音高と演奏情報を別々に入力・統合する研究があり [18]，有効性が示されている．

6 おわりに

本稿では，歌唱を効率的に録音し，音の三要素に分解してそれをインタラクティブに統合する VocaRefiner を提案した．録音では，歌声と音素の自動アラインメントにより，その統合が効率化された．VocaRefiner によって，歌唱力，歌声合成パラメータ調整や歌声編集といった従来の歌声生成のスキルに加えて，インタラクションによる新しい歌声生成スキルが切り拓かれる可能性がある．

「歌声の作り方」のイメージが変わり，分解した状態で要素を選択・編集できることを前提に歌作りするようになる可能性もある．例えば，歌唱として完璧には歌えない人でも，要素に分解することで，全体的な完璧さを求める場合より敷居が低くなる．

今後は，再合成音の品質向上や，よりインテリジェントな機能の付与に取り組んでいきたい．例えば，再生バーや歌詞による頭出しに加え，Songle [17] のような楽曲構造の可視化を伴って録音できたり，背景音楽のキーに応じて，自動的に音高を補正したりといった機能である．

謝辞

本研究では，RWC 研究用音楽データベース（ポピュラー音楽 RWC-MDB-P-2001）を使用しました．本研究の一部は，科学技術振興機構 OngaCREST プロジェクトによる支援を受けました．

参考文献

[1] J. Bonada and S. Xavier. Synthesis of the Singing Voice by Performance Sampling and Spectral Models. *IEEE Signal Processing Magazine*, 24 (2):67–79, 2007.

[2] V. Digalakis and L. Neumeyer. Speaker adaptation using combined transformation and Bayesian methods. *IEEE Trans. Speech and Audio Processing*, 4(4):294–300, 1996.

[3] S. Fukayama, K. Nakatsuma, S. Sako, T. Nishimoto, and S. Sagayama. Automatic Song Composition from the Lyrics exploiting Prosody of the Japanese Language. In *Proc. SMC 2010*, pp. 299–302, 2010.

[4] H. Kenmochi and H. Ohshita. VOCALOID – Commercial Singing Synthesizer based on Sample Concatenation. In *Proc. Interspeech 2007*, 2007.

[5] K. Saino, M. Tachibana, and H. Kenmochi. Temporally Variable Multi-Aspect Auditory Morphing Enabling Extrapolation without Objective and Perceptual Breakdown. In *Proc. ICASSP 2009*, pp. 3905–3908, 2009.

[6] 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. In *Proc. WASPAA 2007*, pp. 215–218, 2007.

[7] F. Villavicencio and J. Bonada. Applying Voice Conversion to Concatenative Singing-Voice Synthesis. In *Proc. Interspeech 2010*, pp. 2162–2165, 2010.

[8] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book*. 2002.

[9] U. Zölzer and X. Amatriain. *DAFX - Digital Audio Effects*. Wiley, 2002.

[10] 河原 英紀, 生駒 太一, 森勢 将雅, 高橋 徹, 豊田 健一, 片寄 晴弘. モーフィングに基づく歌唱デザインインタフェースの提案と初期検討. 情報処理学会論文誌, 48(12):3637–3648, 2007.

[11] 齋藤 毅, 後藤 真孝, 鵜木 祐史, 赤木 正人. SingBySpeaking: 歌声知覚に重要な音響特徴を制御して話声を歌声に変換するシステム. 情報処理学会研究報告 音楽情報科学 2008-MUS-74-5, pp. 25–32, 2008.

[12] 大浦 圭一郎, 間瀬 絢美, 山田 知彦, 徳田 恵一, 後藤 真孝. Sinsy:「あの人に歌ってほしい」をかなえる HMM 歌声合成システム. 音楽情報科学研究会 研究報告 2010-MUS-86, pp. 1–8, 2010.

[13] 藤原 弘将, 後藤 真孝. 混合音中の歌声スペクトル包絡推定に基づく歌声の声質変換手法. 情報処理学会研究報告 音楽情報科学 2010-MUS-86-7, pp. 1–10, 2010.

[14] 酒向 慎司, 宮島 千代美, 徳田 恵一, 北村 正. 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, 45(7):719–727, 2004.

[15] 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. 情報処理学会誌, 53(5):466–471, 2012.

[16] 後藤 真孝, 伊藤 克亘, 速水 悟. 自然発話中の有声休止箇所のリアルタイム検出システム. 電子情報通信学会論文誌 D-II, J83-D-II(11):2330–2340, 2000.

[17] 後藤 真孝, 吉井 和佳, 藤原 弘将, M. Mauch, 中野 倫靖. Songle: ユーザが誤り訂正により貢献可能な能動的音楽鑑賞サービス. 情報処理学会 インタラクション 2012 論文集, pp. 1–8, 2012.

[18] 大島 千佳, 西本 一志, 宮川 洋平, 白崎 隆史. 音楽表情を担う要素と音高の分割入力による容易な MIDI シーケンスデータ作成システム. 情報処理学会論文誌, 44(7):1778–1790, 2003.

[19] 河原 達也, 住吉 貴志, 李 晃伸, 坂野 秀樹, 武田 一哉, 三村 正人, 伊藤 克亘, 伊藤彰則, 鹿野 清宏. 連続音声認識コンソーシアム 2002 年度版ソフトウェアの概要. 情報処理学会研究報告音声言語情報処理 2001-SLP-48-1, pp. 1–6, 2003.

[20] 川上 裕司, 坂野 秀樹, 板倉 文忠. 声道断面積関数を用いた GMM に基づく歌唱音声の声質変換. 電子情報通信学会技術報 音声（SP2010-81）, pp. 71–76, 2010.

[21] 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, 52(12):3853–3867, 2011.

[22] 中野 倫靖, 後藤 真孝. 歌声・音声分析合成のための F0 適応多重フレーム統合分析に基づくスペクトル包絡と群遅延の推定法. 情報処理学会 音楽情報科学研究会 研究報告 2012-MUS-96-7, pp. 1–9, 2012.

[23] 中野 皓太, 森勢 将雅, 西浦 敬信, 山下 洋一. 基本周波数の転写に基づく実時間歌唱制御システムの実現を目的とした高品質ボコーダ STRAIGHT の高速化. 電子情報通信学会論文誌, 95-A(7):563–572, 2012.

未来ビジョン

本研究では，インタラクションに基づく「未来の歌手」を考えた．現在「神業的なパラメータ調整」による歌声合成,「精緻にコントロール」された人間の歌唱,「予想外に曲にマッチした新しい」声質の編集を行う人達は，全て歌声を生成する人であり，広義の歌手といえる．

このような未来では，歌手は三種類の方法で，その能力を拡張することが可能である．第一に様々な経験によって，制作者自身の能力を拡張していくことができる．第二に人間のパートナーを持ち，お互いに自分にはないスキルを発揮することで，歌声合成結果の質を上げることも考えられる．そして第三に計算機をパートナーとして，インタラクションによって音楽制作の質を上げていける可能性がある．

計算機で音楽制作をすると，いつでも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追求して表出できる利点がある．現時点の計算機は，本稿で述べたように限界もあるが，将来的には,「この歌い方はどうか？」「ここはもう少し声を張り上げて歌うと感動するのでは？」などといった提案を計算機側からできれば，手軽さと気軽さが，より高品質な音楽制作につながる可能性がある．

例えば，計算機からの提案によって，ユーザの想像力が刺激されれば，新しい表現が生まれるかもしれない．すなわち，ユーザの視野が広がり，能力の拡張につながる可能性がある．そして，さらにその表現を受けて，計算機が新たな提案をする．そのような人間と計算機のインタラクションと，能力の拡張，表現力の拡大が，著者が目指す歌声生成の未来である．