情報処理学会研究報告, 音楽情報科学, Vol.2010-MUS-86, No.3, pp.1-10, July 2010.
IPSJ SIG Technical Report

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

中野倫靖[†1]　後藤真孝[†1]

本稿では、ユーザの歌唱音声からその声色（こわいろ）変化を真似て歌声合成するシステム VocaListener2 を提案する。我々が以前開発した VocaListener では、音高と音量のみを真似て歌声合成パラメータを推定していたが、VocaListener2 ではそれを拡張して声色変化にも対応する。従来、主に声質変換やモーフィングのために、声質を操作する技術はあったが、ユーザ歌唱の声色変化を反映することはできなかった。VocaListener2 を実現するために、まず VocaListener によってユーザ歌唱の音高と音量を真似た多様な歌声を合成して声色空間を構成し、その結果を用いてユーザ歌唱の声色変化を反映して合成する。市販の歌声合成システムを用いて実験した結果、VocaListener2 では音高と音量に加えて声色変化も真似ることができていた。

VocaListener2: A Singing Synthesis System Mimicking Voice Timbre Changes in Addition to Pitch and Dynamics of User's Singing

Tomoyasu Nakano[†1] and Masataka Goto[†1]

In this paper, we propose a singing synthesis system, *VocaListener2*, that automatically synthesizes a singing voice by mimicking timbre changes of a user's singing voice. The system extends our previous system called *VocaListener* that can estimate singing synthesis parameters of only pitch (F_0) and dynamics (power) from the user's singing voice. Although most previous techniques for manipulating voice timbre have focused on voice conversion and voice morphing, they cannot deal with the timbre changes during singing. To develop VocaListener2, we first construct a voice timbre space on the basis of various singing voices that mimic the pitch and dynamics of the user's singing voice by using the VocaListener. In this space, the timbre changes can be reflected to the synthesized singing voice. In our experiences with singing synthesis systems on the market, we found the timbre changes as well as the pitch and dynamics can be mimicked.

1. はじめに

本研究では、歌声合成システムを利用する多様なユーザが、魅力的な歌声を自由自在に合成して楽曲等を制作し、歌唱という音楽表現の可能性を広げることを支援できる技術の開発を目指す。人間のような歌声を人工的に生成できる歌声合成システムは、多様な歌声での合成が容易に行え、歌唱の表現を再現性高くコントロールできることから、歌唱付き楽曲の制作における可能性を広げる重要なツールである。2007 年以降、市販の歌声合成ソフトウェアを使った楽曲制作を楽しむユーザが急増し、その利用拡大に対する社会的関心の高さからさまざまなメディアに取り上げられてきた。内閣府による海外向け広報誌においても紹介されている[1] ように、歌声合成ソフトウェアを用いた楽曲が動画共有サービス等に多数投稿され、制作しているユーザが増えただけでなく、そうした楽曲を楽しむリスナーも増えた。また一方で、そうして創られた作品は、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 を介した音楽の共同制作や新しいコミュニケーションを生み出している現状がある[2),3)]。さらに、高品質な歌声合成技術の実現を目指すことは、人間の歌声知覚・生成機構の解明にも繋がる取り組みである。

我々は以前、入力としてユーザが歌唱音声を与え、その歌唱の音高と音量を真似るように、既存の歌声合成ソフトウェアの合成パラメータを調整できるシステム VocaListener を開発した[4),5)]。本稿ではそれを拡張し、ユーザ歌唱の声色（こわいろ）変化を歌声合成結果に反映できる VocaListener2 を提案する。そこで、これまでの VocaListener を VocaListener1 として本提案と区別する。本研究が目指すものは、細かなパラメータ操作なしに表情豊かな歌声を合成でき、合成された歌声によってどのような表現をしたいのか、どのようなメッセージを伝えたいのかに、より注力して歌声を合成できることである。しかし、これまでの VocaListener1 は音高と音量しか扱えず、ユーザ歌唱の表情や歌い方を表現しきれていなかった。ユーザ歌唱中の声色変化を真似て、歌詞やメロディーに合わせて歌声合成結果に反映できれば、より魅力的な歌声合成の実現につながると考えられる。

従来、声質として、歌声や話声の声質変換や声質モーフィングに関する研究がなされてきた。歌声合成においては、ユーザによる手作業（マウス）での数値パラメータ調整[6)]、二人

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

の歌唱者による同一歌詞の歌唱音声からの声質のモーフィング[7] や、感情を変えて歌った同一歌唱者の複数の歌唱へ応用した感情モーフィング[8] がある。また話声合成においては、異なる話者間の声質変換や[9),10)]、感情音声合成に関する研究があった[11)–18)]。感情音声合成に関しては、話声の韻律や話速を扱うものが多いが、感情変化に伴なう声質変換[11),16)–18)]も研究されている。また、話声のモーフィングに関して、複数音声からの平均声生成[19] や、複数音声から比率を推定してユーザ音声に近い声にモーフィングする[20] 研究もある。

しかし、こうした従来の研究のほとんどは、異なる音声間での変換やモーフィングを対象とし、本研究が対象とするような歌唱中の変化を操作することはできなかった。ここで「声質」という用語は、個人を特定できる音響的な特性や聴覚上の違いだけでなく、異なる発声様式によって生じる声の違い（唸り声、囁き声等）や、明るい声や暗い声といった聴感上の印象（表現語）の違いなど、多様な意味合いで使われているため、本研究ではそういった歌唱中の変化を表す際、声質という単語と区別して「声色変化」という単語を用いる。

従来、このような声色変化をユーザが明示的に扱える技術には、歌声合成システム Vocaloid[6] があった。Vocaloid では、複数の数値パラメータ[⋆1]を各時刻で調整することで、歌唱音声のスペクトルを操作して声色変化を伴った歌声合成が実現できる。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、ほとんどのユーザはこれらを変更しないか、変更するにしても曲毎に一括で変更したり、大まかに変更したりしていた。

以上の問題を解決するために、本研究では VocaListener1 によりユーザ歌唱と同一歌詞で、音高と音量を真似た複数の多様な歌声を合成し、それらの歌声全てから声色変化に寄与する成分を表す空間（声色空間）を構成する。そして、その空間上でのユーザの声色変化を反映させて歌声合成する。また、ユーザ歌唱を真似るだけでは、歌唱によるユーザの表現力の限界を超えられないため、声色変化を調整できるインタフェースについても提案する。

これ以降、**2** 章で我々が以前開発した VocaListener1 の問題点を提起し、**3** 章でそれを解決する VocaListener2 の実現方法について述べた後、**4** 章で実験を行う。最後に、**5** 章で本技術の応用について議論し、**6** 章で本研究の意義と今後の課題について述べる。

2. VocaListener1 の機能とその問題点

本章では VocaListener1 の概説と、VocaListener2 を実現するための課題について述べる。

2.1 VocaListener1: ユーザ歌唱を真似る歌声合成パラメータ推定システム[4),5)]

VocaListener1 は、既存の歌声合成ソフトウェア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 1**）。パラメータの反復推定により、推定精度が従来研究[21] 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、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。

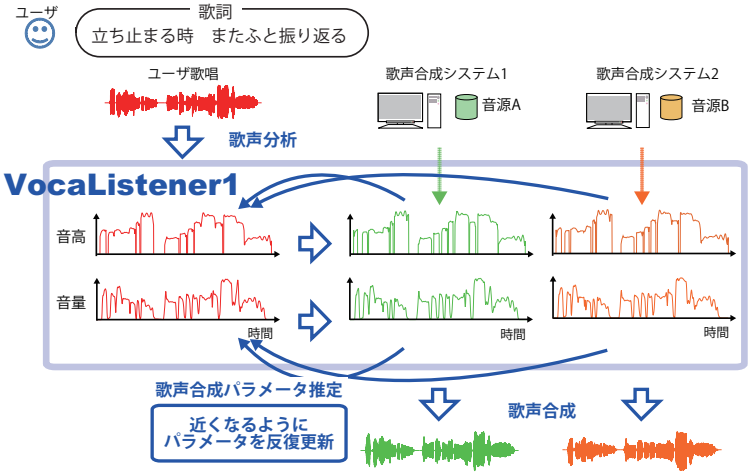

図 1　VocaListener1 によるユーザの歌声とその歌詞からの歌声合成パラメータ推定の概要および推定結果

図に示したように VocaListener1 では、ユーザ歌唱を音高と音量に関して真似して歌声合成できる。具体的な数値に関しては文献 5) に述べられている。また合成結果の具体例は、ホームページ[⋆2]や動画コミュニケーションサービス『ニコニコ動画』[⋆3]上で視聴できる。

2.2 ユーザ歌唱の声色変化を真似る歌声合成の実現方針

声色変化を対象として「ユーザ歌唱を真似る」ためには、前節で説明した VocaListener1

⋆1 音高・音量以外では、Vocaloid1 では Note Velocity, Resonance, Harmonics, Noise, Brightness, Clearness, Gender Factor, Vocaloid2 では VEL, BRE, BRI, CLE, OPE, GEN がある。

⋆2 http://staff.aist.go.jp/t.nakano/VocaListener/index-j.html
⋆3 http://www.nicovideo.jp/mylist/7012071/

と同様、既存の歌声合成システムにおける声質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。しかし結論からいうと、この方法は実現可能性はあっても、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、声質や声色変化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、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[6] では、操作できるパラメータが一部異なる。したがって、声質パラメータ毎に最適化した方法を仮に実現しても、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。

一方、クリプトン・フューチャー・メディア株式会社の応用商品である「初音ミク・アペンド (MIKU Append)[⋆1]」は、「初音ミク[⋆2]」と同一歌唱者の声で、DARK, LIGHT, SOFT, SOLID, SWEET, VIVID の 6 種類の声色で歌声合成できる。しかし、これらの音源をフレーズ毎に切り替えながら合成することはできても、歌声合成システム上でこれらの中間の状態を作り出すことは困難であり、例えば「LIGHT と SOLID の中間の声」で歌い始めた後、徐々に「初音ミクの声」に切り替わる、といった滑らかな変化を実現するのは難しい。

したがって、これらの問題を解決するには、歌声合成システム内のパラメータ操作だけでは不十分で、外部の信号処理が必要となる。そこで、まず VocaListener1 で音高と音量を真似て合成した後、その合成歌唱を利用しながら、声色変化を信号処理で反映する。

2.3 ユーザ歌唱の声色変化を真似る歌声合成の実現課題

ユーザ歌唱の声色変化を真似る歌声合成を実現するためには、「声色変化」を「真似る」という問題を解決する必要があり、具体的な実現課題は以下の二つである。

実現課題 (1): 声色変化をどのように表現するのか。

実現課題 (2): ユーザ歌唱の声色変化をどのように反映させるのか。

ここで声色の違いとは、本稿では、前節の説明における初音ミクと初音ミク・アペンドの違いに相当し、それはスペクトル包絡の形状の違いとして定義できる（**図 2**）。しかしスペクトル包絡形状の違いには、図 2 に示すように、音韻の違いや個人性の違いも含まれる。したがって、そのような成分を抑制した時間変化が声色変化といえる。そして、そのような声色変化を反映したスペクトル包絡の時間系列を新たに生成できれば、ユーザ歌唱の声色変化を真似た歌声合成が実現できる。

⋆1 http://www.crypton.co.jp/cv01a/
⋆2 http://www.crypton.co.jp/mp/pages/prod/vocaloid/cv01.jsp

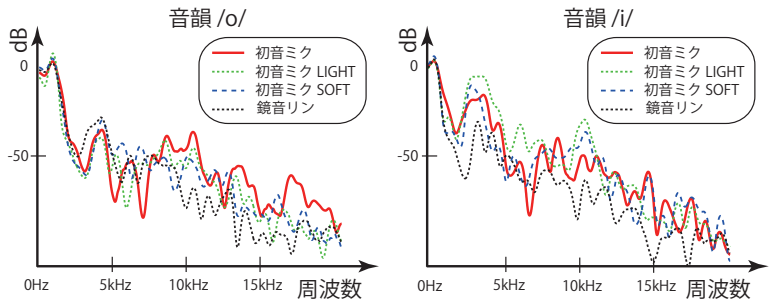

図 2 音韻の違いと歌唱者の違いによるスペクトル包絡形状の違いの具体例。

3. VocaListener2: ユーザ歌唱の声色変化を真似る歌声合成システム

本章では、前章で述べた課題の解決法を述べ、VocaListener2 の実現方法を説明する。

3.1 ユーザ歌唱の声色変化を真似る歌声合成の課題における解決方針

2.3 節で述べた実現課題 (1) を解決するために、まず VocaListener1 を用いて、ユーザ歌唱を真似て、時刻が同期した複数の歌唱者による歌唱音声を自動的に生成する。ここで、合成対象となる同一歌唱者の声色が異なる歌唱（例: 初音ミクと初音ミク・アペンド）も同時に合成する。これによって、各時刻において音高・音量・音韻が同期した歌唱が得られるため、これら全てを活用して、声色変化以外の成分を抑制した**声色空間**を構成する。そこでは、全ての歌唱が各時刻において声色空間上の一点に対応し、その時間変化は、声色空間上の時間変化する軌跡として表現できる。

続いて、実現課題 (2) を解決するために、VocaListener1 による同一歌唱者の声色が異なる合成結果（同期した歌唱）の、声色空間上における複数の軌跡について、それらを含むような多面体（ポリトープ）とその時間軌跡を考え、これを**声色変化チューブ**と呼ぶ。声色空間を M 次元空間とすると、合成対象の声色は、各時刻 t において J 個の M 次元ベクトル $\mathbf{z}_{j=1,2,\ldots,J}(t)$ がその空間上に存在し[⋆3]、これら J 個の点 $\mathbf{z}_j(t)$ に囲まれた内側が、合成したい同一の歌唱者の変形可能な領域と本研究では仮定する。つまり、この時々刻々と変化する多面体（M 次元ポリトープ）が声色変化可能な領域であると考える。したがって、同じく声色空間の別の場所に存在するユーザ歌唱の軌跡 $\mathbf{u}(t)$ を、声色変化チューブ内になるべ

⋆3 初音ミクと初音ミク・アペンドの 7 種類の声色を考えるなら $J = 7$ である。

く入るようにシフト・スケーリングさせた $\mathbf{u}'(t)$ を得ることで、各時刻における声色空間上の合成目標位置を決定する。その位置から出力する合成歌唱のスペクトル包絡を生成することで VocaListener2 を実現する。

3.2 VocaListener2 の処理概要

処理の流れを**図 3** に示す。VocaListener2 では、入力としてユーザの歌唱音声を与え、出力としてユーザの声色変化を反映して、特定の歌声 Z で真似た合成歌唱を得る。図中、Z_1〜Z_4 は初音ミクと初音ミク・アペンドに相当する。まず、ユーザ歌唱から VocaListener1 を用いて、その歌い方を真似た歌声を複数生成する（A）。これによって、ユーザ歌唱と時刻が同期した複数の多様な歌唱音声が得られる。

続いて、それぞれの歌唱音声を分析し、音高（F_0）による影響を除去したスペクトル包絡を推定する（B）。ここでスペクトル包絡を推定するのは、2.3 節で述べたようにそれが声色変化を反映するからであり、F_0 の影響を除去するのは、入力歌唱には男女の違いなどによる F_0 の絶対値の違いが存在するからである。そのようにして得たスペクトル包絡に基づき、声色を反映した M 次元の声色空間を構成する（C）。最後に、声色空間上のユーザ歌唱の軌跡を Z_1〜Z_4 によって構成される声色変化チューブによく収まるように、シフトとスケーリング操作を行ない（D）、ユーザ歌唱の声色変化を反映して歌声合成する（E）。

以降、それぞれの処理について具体的な実現方法を説明する。

3.3 歌声分析: 歌唱音声からのスペクトル包絡系列の推定（図中 B に相当）

声色変化をよく表す音響的な特性として、本研究ではスペクトル包絡を対象とし、歌唱音声から推定する。ここで、それぞれの歌唱の F_0 の影響を除去してスペクトル包絡を得るために、音声分析合成系 STRAIGHT[22] を用いる。このスペクトル包絡（STRAIGHT スペクトルと呼ばれる）に基づいて処理を行うのは、それを変形して高品質な再合成が行えることが知られているからである[7]。

3.4 歌声分析: 声色空間の構成（図中 C に相当）

スペクトル包絡の時間系列から声色変化に寄与する成分以外を、部分空間法に基づいた処理によって抑制して声色空間を構成する。**図 4** に処理の概要を示す。

VocaListener1 を用いて複数の歌唱を合成したことで、あるフレーム時刻における全歌唱者のスペクトル包絡は、個人性（声質）や声色の違いに相当する変動のみが存在すると考えられる。これは、音高・音量・音韻が同一となるように VocaListener1 によって真似ているからである。ここで、男女の違い等による絶対的な音高の違いは存在するが、音高の違いは STRAIGHT による包絡推定によって除去されていると仮定する[⋆1]。したがって、フレーム毎に主成分分析を行った結果、フレーム毎の異なる声色を持つ歌唱間で分散が大きい低次元の部分空間は、声色変化の寄与が大きな空間として考えることができる[⋆2]。

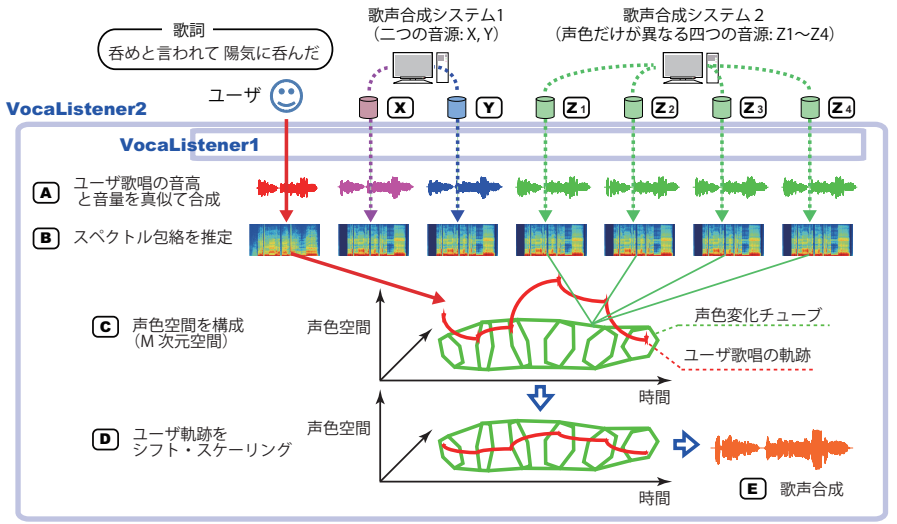

図 3 VocaListener2 における処理の流れ。ユーザ歌唱の歌い方を真似た複数の声質と複数の声色の歌唱から、声色空間を構築し、ユーザ歌唱の声色変化を真似るように合成する。

部分空間法に基づいたこのような方法は、音韻性と話者性の分離に基づいた話者認識[23] や声質変換[24] において有効性が確認されている。従来研究[23),24)] では、話者毎に部分空間を構成することで、音韻性（低次部分空間: 変動が大きな成分）と話者性（高次部分空間: 変動が小さな成分）を分離していたが、本研究ではそれをフレーム毎に行う。しかしそのままでは、各フレームで異なる空間が構成されることになり、全フレームを統一的に扱えない。

そこで、まずはフレーム毎の部分空間における低次 N 次元のみを保存して、元の空間に戻すことで、声質・声色変化に寄与する成分以外を抑制する。続いて、全歌唱の全フレームを用いて一度に主成分分析を行い、その低次 M 次元の空間を声色空間として扱う。このよ

⋆1 実際には、F_0 が大きく異なると、スペクトル包絡の形状にも異なる可能性があるが、数半音の違いの音は STRAIGHT によって吸収できると仮定する。また、それ以上の音高の違いによるスペクトル包絡の違いは、声色の違いとして扱われることになる。
⋆2 個人性も残ると考えられる。

うな処理によって、異なる歌唱者の全てのフレームが同じ空間上で扱えるだけでなく、音韻などの文脈に伴なう声色変化に関係する成分を、低次元で効率的に表現できる[*1]。さらに、このような処理による余計な成分の抑制は、次節で述べるユーザ歌唱との対応付けにおいても重要と考えられる。

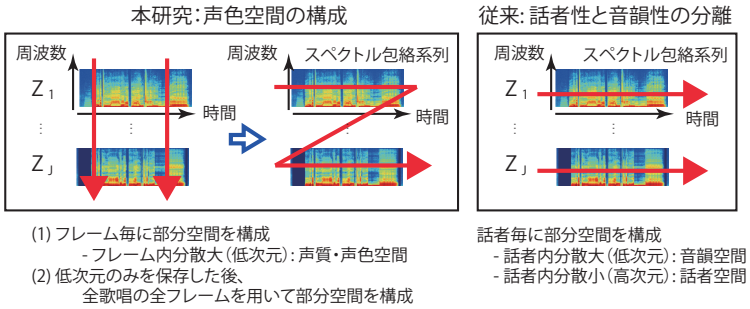

図 4 部分空間法による声色空間の構築の概要と、従来研究[23),24)] との違い。

3.5 歌声分析: 声色空間におけるユーザ歌唱との対応付け（図中 D に相当）

声色空間上のユーザ歌唱の軌跡が、声色変化チューブ内にできるだけ存在するように、ユーザ軌跡を対応付ける。この操作を行うことで、ユーザ歌唱の声色変化を反映させることができる。このような対応付け方法には、様々な方法が考えられるが、本稿では単純な方法として、ユーザ歌唱と声色変化チューブのそれぞれにおいて、各次元で 0〜1 の値となるようにシフト・スケーリングを行った。

3.6 歌声合成: 声色空間上の軌跡からの歌声合成（図中 E に相当）

声色空間上の一点から、それに対応付くようなスペクトル包絡を生成する。ここで、実際の声色空間上での各声色として、ある時刻における初音ミクと初音ミク・アペンド（DARK, LIGHT, SOFT, SOLID, SWEET, VIVID）の配置を**図 5** に示す（ただし、左上の声色変化チューブはイメージ図である）。図に示すように、それぞれの点にはスペクトル包絡が対応付いており、これに基づいて、ユーザ歌唱の声色変化を反映させるスペクトル包絡を生成することが課題である。

従来、二人もしくは複数話者のスペクトル包絡間のモーフィングでは、周波数軸方向に特

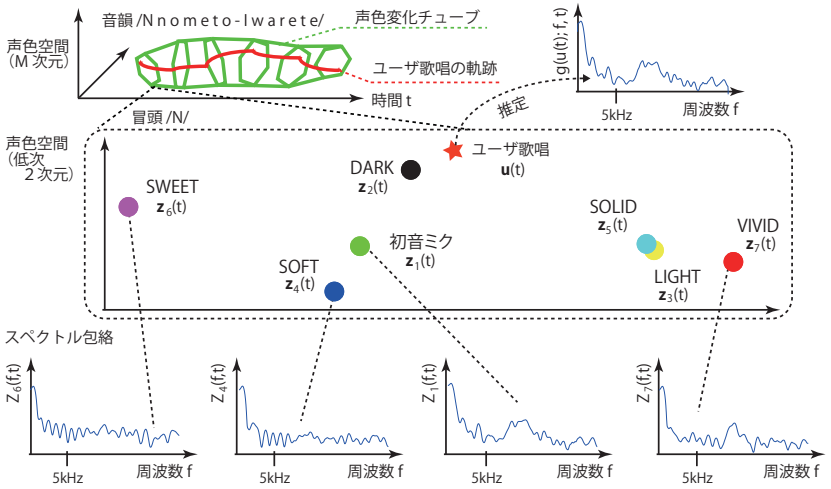

図 5 声色空間における各点には、スペクトル包絡がそれぞれ対応付いており、ユーザ歌唱がその点に重なった時は同じスペクトル包絡が生成されるように補間する。ここでは、実際に構成した声色空間上のある時刻における初音ミクと初音ミク・アペンド（DARK, LIGHT, SOFT, SOLID, SWEET, VIVID）の配置を示す。

徴点を適切に設定して非線形に伸縮させることで、品質を保ったモーフィングが行えることが知られている[7),19)]。しかしここでは、同一歌唱者・同一音高のモーフィングに相当するため、そのような非線形伸縮をすることなく、スペクトル包絡の各周波数毎の強調・抑制処理のみで声色変換が可能であると仮定する。

そこで、スペクトル包絡をそのまま使うのではなく、標準的な声（例えば初音ミク・アペンドでない初音ミク）を基準としてそこからの変形比率として表し、この比率をまずフレーム毎に推定する。本稿では、これを**スペクトル変形曲線**と呼び、全時刻のスペクトル変形曲線を合わせて**スペクトル変形曲面**と呼ぶ。ここで、ユーザ歌唱が声色空間上で各声色の点と重なりあった場合には、それと同じスペクトル変形曲線を生成する制約を満たすように推定する。そのために、Radial Basis Function を用いた Variational Interpolation[25)] を応用して適用する。ここで、時刻 t、周波数 f における各声色のスペクトル包絡を $Z_{j=1,2,\ldots,J}(f,t)$、その $Z_1(f,t)$ に対するスペクトル変形曲面を $Zr_j(f,t)$ とし、声色空間上でのユーザ歌唱を $\mathbf{u}(t)$、各声色を $\mathbf{z}_j(t)$ とすると、次の制約付きの方程式を解くことで、声色空間上でのユーザの声色を真似るためのスペクトル変形曲線を得る。

⋆1 表現力の高い空間を得るために、声色空間を構成する際に用いる歌唱者は多い方が望ましい。

$$Zr_j(f,t) = \log\left(\frac{Z_j(f,t)}{Z_1(f,t)}\right) \tag{1}$$

$$g(\mathbf{u}(t);f,t) = \sum_{k=1}^{J}\left(w_k(f,t)\cdot\phi\left(\mathbf{u}(t)-\mathbf{z}_k(t)\right)\right) + P(\mathbf{u}(t);f,t) \tag{2}$$

$$Zr_j(f,t) = \sum_{k=1}^{J}\left(w_k(f,t)\cdot\phi\left(\mathbf{z}_j(t)-\mathbf{z}_k(t)\right)\right) + P(\mathbf{z}_j(t);f,t) \tag{3}$$

$$g(\mathbf{z}_j(t);f,t) = Zr_j(f,t) \tag{4}$$

$$P(\mathbf{x};f,t) = p_0(f,t) + \sum_{m=1}^{M} p_m(f,t)\cdot x^{(m)} \tag{5}$$

ここで $Zr_i(f,t)$ は式 (1) のように対数をとり、比率を対数軸上に線形に変換させることと、推定結果が負の値を取ることを許容する。また w_j が混合比率であり、$P(\cdot)$ は式 (5) のように、ベクトル $\mathbf{x}$ として $\mathbf{z}_j(t)$ もしくは $\mathbf{u}(t)$ を変数とする M 変数一次多項式（係数が $p_{m=0,\ldots,M}$）である。$\phi(\cdot)$ は、ベクトル間の距離を表す関数であり、本稿では $\phi(\cdot)=|\cdot|$ とする[⋆1]。式 (4) が前述の制約に相当し、声色空間を $M=3$ 次元とすると以下の行列で書ける。

$$\begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1J} & 1 & z_1^{(1)} & z_1^{(2)} & z_1^{(3)} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2J} & 1 & z_2^{(1)} & z_2^{(2)} & z_2^{(3)} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi_{J1} & \phi_{J2} & \cdots & \phi_{JJ} & 1 & z_J^{(1)} & z_J^{(2)} & z_J^{(3)} \\ 1 & 1 & \cdots & 1 & 0 & 0 & 0 & 0 \\ z_1^{(1)} & z_2^{(1)} & \cdots & z_J^{(1)} & 0 & 0 & 0 & 0 \\ z_1^{(2)} & z_2^{(2)} & \cdots & z_J^{(2)} & 0 & 0 & 0 & 0 \\ z_1^{(3)} & z_2^{(3)} & \cdots & z_J^{(3)} & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_J \\ p_0 \\ p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} Zr_1 \\ Zr_2 \\ \vdots \\ Zr_J \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{6}$$

ここで ϕ_{ij} は $\phi(\mathbf{z}_i(t)-\mathbf{z}_j(t))$ を表し、(f,t) や (t) は省略して記述した。

このようにして推定された w_j と p_m を用いて、式 (2) によってスペクトル変形曲面を生成する。続いて、合成の不自然さを減らすために、フレーム毎に上限と下限を定めて、声色変化チューブ外にユーザ歌唱が存在した場合の影響を低減する。また、時間–周波数平面上の平滑化処理により、急峻すぎる変化を低減してスペクトルの連続性を保つ。最後に、基準とした歌唱音声のスペクトル包絡をこのスペクトル変形曲面を用いて変形し、それをSTRAIGHT で合成することでユーザ歌唱の声色変化を真似た合成歌唱を得る。

⋆1 その他、$\phi(\cdot)=|\cdot|^2\log(\cdot)$ や $\phi(\cdot)=|\cdot|^3$ 等が使われることがある。

3.7 インタフェース構築: ユーザによる声色変化の調整機能

以上のような処理により、ユーザ歌唱の声色変化を真似た歌声合成が実現できるが、ユーザ歌唱を真似るだけでは、歌唱によるユーザの表現力の限界を超えることができない。そこで、表現の幅を拡げるため、推定結果に基づいて声色変化を操作できるインタフェースを提案する。そのようなインタフェースでは、以下の三つの機能を持つ。

(1) 声色変化のスケールを変えて声色変化の度合いを変更する機能
スケールを大きくして抑揚ある歌声を合成したり、逆にスケールを小さく声色変化を抑えたりして合成できる。

(2) 声色変化をシフトして声色変化の中心を変更する機能
声色変化の中心を変えることで、それぞれの声色を中心とした声色変化に変換できる。

(3) 声色変化を部分的にシフト・スケーリングして微調整する機能
上記二つの機能を部分的に適用することで、細かな修正を可能とする。

4. 実　　験

本章では、合成歌唱を得る過程での、各処理における妥当性を検証する。

4.1 実験条件

本章で示す実験は、RWC 研究用音楽データベース（音楽ジャンル）RWC-MDB-G-2001[26] No.91「大漁船」を用いて行った。これ以降、歌唱音声信号はサンプリング周波数 44.1kHz のモノラル音声信号を扱い、処理の時間単位は 1 msec とする。

声色空間を構成するために利用する歌声合成システムとしては、Vocaloid と Vocaloid2[6] を採用し、その応用商品として現在市販されている歌声合成ソフトウェアのうち、日本語歌唱を合成できる全 17 種類を用いた（音高と音量以外のパラメータ全てにデフォルト値を用いた）。そのうち男性歌唱が 3 種類[⋆2]、男性歌唱に対して 1 オクターブ上げて合成した女性歌唱が 14 種類[⋆3]である。ここで、初音ミクと初音ミク・アペンドの 7 種類を合成対象の歌声とし、それらから声色変化チューブを構成した。

それぞれの歌唱音声を STRAIGHT によって、各時刻でスペクトル包絡を推定する際に

⋆2 KAITO (Vocaloid1)、がくっぽいど、氷山キヨテル (以上、Vocaloid2)。

⋆3 MEIKO (Vocaloid1)、初音ミク、鏡音リン、鏡音レン、巡音ルカ、初音ミク・アペンド（6 種類）、メグッポイド、歌愛ユキ、SF-A2 開発コード miki（以上、Vocaloid 2)。鏡音レンは男性とされているが、鏡音リンと同一歌唱者であるため、女性として扱った。

表 1　フレーム毎の 17 種類の歌唱音声に対する主成分分析における累積寄与率と次元数 N の関係

累積寄与率 R [%]	50	55	60	65	70	75	80	85	90	95
次元数（平均）	1.29	1.62	1.97	2.40	2.89	3.48	4.18	5.04	6.16	7.70
次元数（標準偏差）	1.01	1.27	1.51	1.84	2.20	2.64	3.14	3.77	4.59	5.73

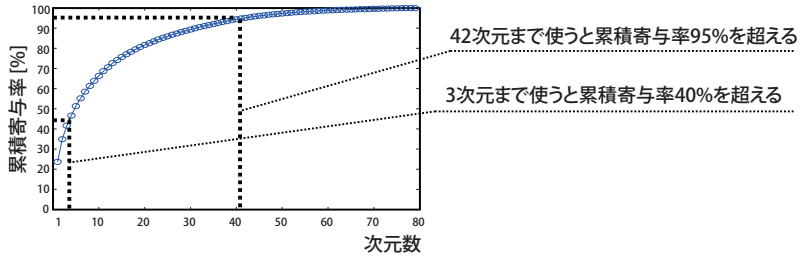

図 6　全歌唱の全フレームを用いた主成分分析における累積寄与率と次元数 M の関係

は、分析フレーム長は F_0 同期とし、各時刻でスペクトル包絡を推定する際の FFT 長は 4096 点とした。声色空間を構成するための主成分分析において、スペクトル包絡をそのまま用いず、離散コサイン変換を行って、0 次（直流成分）を除いた低次 80 次元のみを用いた。低次 80 次元あれば、次元数を落としながら、STRAIGHT スペクトルを良く再現できたためである。その後、フレーム毎の異なる声色を持つ歌唱間での主成分分析では累積寄与率 80%を超える次元数を用い（次元数 N はフレーム毎に可変）、全歌唱の全フレームを用いた主成分分析では上位 3 次元（$M = 3$）を用いて声色空間を構成した。またこれらの処理は、全ての歌唱で F_0 が存在する有声区間のみを用いた。

4.2　実験 A: 部分空間法によって構成された声色空間の特性の確認

本実験では、便宜上のユーザ歌唱として「大漁船」の無伴奏の男性歌唱（55 sec）を用いて、部分空間法による声色空間の構成に関してその特性を確認する。

声色空間の構成において、フレーム毎に 17 種類の歌唱音声に対して主成分分析を行った際に、累積寄与率 R%を超える次元数 N を求め、その N の全フレームでの平均と標準偏差を**表 1** に示す。全歌唱の全フレームを用いた主成分分析を行った場合の、累積寄与率と次元数 M の関係を**図 6** に示す。また、声色空間上の低次 2 次元における各声色の時間変化の一例と全フレームにおける平均ベクトルとその分布を**図 7** に示す。

表 1 からは、スペクトル包絡の形状に関して、平均 7.7 次元程度あれば十分良く表現できることが分かる。声色変化だけに着目したい場合は、より少ない次元数で表現できる可能性があるといえる。また、累積寄与率とともに次元数の標準偏差が大きくなるが、これは声色変化チューブが細くなったり太くなったりすることを意味し、歌詞やメロディーによっては、各声色がそれぞれ近い位置に配置される状況があるといえる。

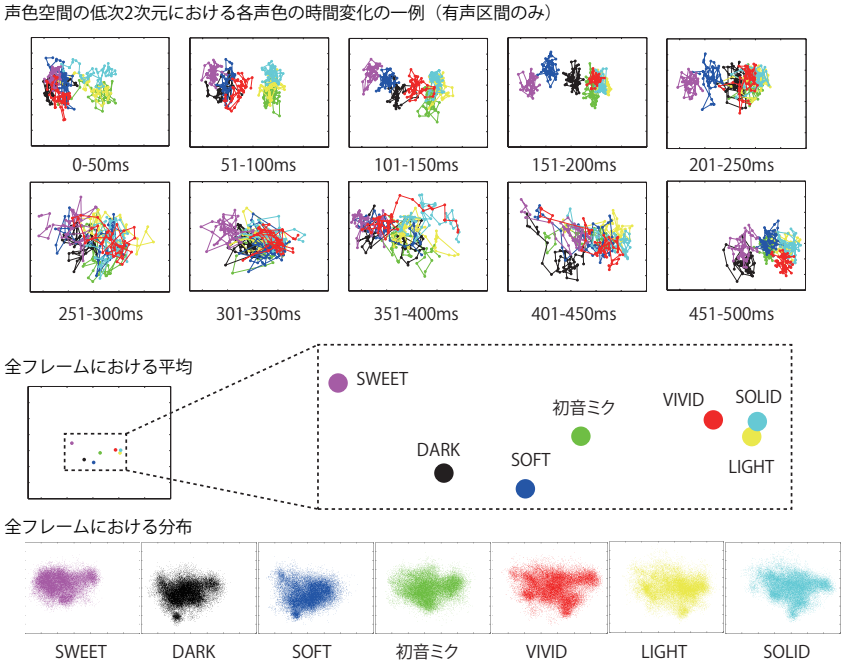

図 7　声色空間の低次 2 次元における各声色の時間変化の一例（上段）、および実験で用いた RWC-MDB-G-2001 No.91「大漁船」の声色毎の全フレーム（55 sec）における平均（中段）とその分布（下段）

また図 6 からは、全歌唱の全フレームを用いた主成分分析では、歌唱や音韻の多様性が含まれるため、フレーム毎の 17 種類の歌唱音声に対して主成分分析を行った場合よりも多くの次元を用いなければ、声色変化を説明できないことが分かる。しかし図 7 中段の結果からは、上位 2 次元に限ってみても、その平均ベクトルは比較的分離されていた。ここで図 7 下段では各声色の分布が重なりあう部分が多く見えるが、実際にはそれぞれが連動して動いていることが、時間変化の一例（上段）から分かる。

以上の結果から、本手法によって得られる声色空間は、フレーム毎に行った短時間の分析

においては 8 次元以下、1 分近い歌唱の全体においても 41 次元以下で、スペクトル包絡の 95%を説明できるといえる。その上で、図 7 の結果を考えると、低次部分空間を活用することで、声色空間を少ない次元で表現できることを示唆している。

最後に、フレーム毎の主成分分析において、低次 1 次元を削除して再合成すると、初音ミクと初音ミク・アペンドの聞こえの差が小さくなることも確認した。したがって低次 1 次元は初音ミクと初音ミク・アペンド間の声色の違いを表していることが示唆される。また、初音ミクと初音ミク・アペンドのそれぞれの合成歌唱の聴取印象は、VIVID と SOLID と LIGHT が類似し、DARK と SOFT が類似しており、SWEET はそれらとは大きく異なるが、各平均ベクトルはそれを反映するような配置となっていた（図 7）。ただし、これらは定性的で主観的な印象であるため、今後のさらなる検証が必要である。

4.3 実験 B: ユーザ歌唱の声色変化を真似る歌声合成結果の確認

本実験では、まず便宜上のユーザ歌唱として「大漁船」の男性歌唱（55 sec）を初音ミク・アペンドで真似た六種類の歌唱について、それを手作業で切り貼りした歌唱音声を対象とする。このような実験によって、パラメータ推定結果の適切性を判断できる。ただしここでは、声色空間を構成した際にも初音ミク・アペンドで真似た合成歌唱が含まれているため、合成時のパラメータを一部変更して VocaListener1 で合成した歌唱も用いる。具体的には、ユーザ歌唱としてそのまま切り貼りした歌唱（Closed 実験）と GEN パラメータを 90 に変更して 2 半音下げた歌唱（Open 実験）の二種類を入力した。

図 8 に、声色空間上におけるそれぞれの声色とのユークリッド距離を示す。Closed 実験の結果からは、正解がほぼ推定できたことから、適切な声色空間が構成され、シフト・スケーリングが適切に行えているといえる。また Open 実験の結果では、Closed 実験の結果よりも結果がばらついていた。ここで、LIGHT と SOLID と VIVID が相互に影響を受けていたり、DARK と SOFT が影響を受けていることがあった。これは、図 7 の結果からも、これらの声色が近くに配置されていることが原因と考えられる。

この結果を用いてスペクトル包絡を生成した結果、入力の声色変化を反映して歌声合成できた。また、入力として「大漁船」の男性歌唱を与えた場合にも、声色変化させながら歌声合成できることを確認した。その際のスペクトル包絡とスペクトル変形曲面を**図 9** に示す。

5. 議　　論

本章では、本研究で提案したシステムのさらなる応用可能性について図 3 の出力部（スペクトル包絡の生成）と入力部（声色変化チューブの構成）の変更について議論する。

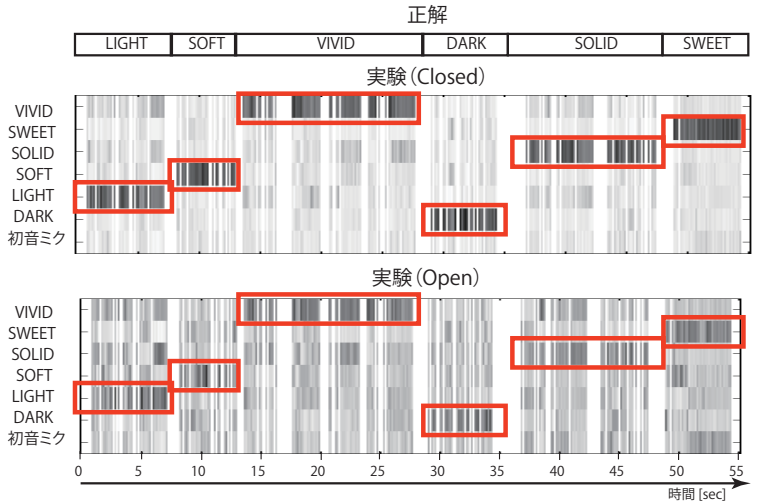

図 8 ユーザ歌唱として初音ミク・アペンド歌唱を切り貼り（LIGHT, DARK, SOLID, SOFT, VIVID, SWEET の順）した歌唱を与えた場合の声色空間上での各声色とのユークリッド距離（濃いほど距離が近い）。上段は全く声色空間を構成する際の初音ミク・アペンドと同じ歌唱を入力として与え、下段は GEN パラメータを変更（GEN=90）して 2 半音下げた歌唱を入力として与えた場合。

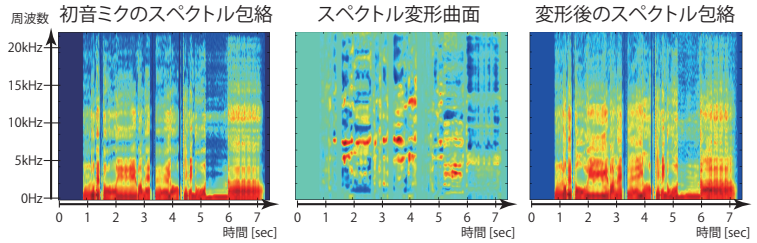

図 9 「大漁船」冒頭 1 フレーズにおける初音ミクのスペクトル包絡（左）、ユーザ歌唱を真似るためのスペクトル変形曲面（中）、最終的に生成されたスペクトル包絡（右）。

5.1 応用 (1): 出力部（スペクトル包絡生成）の変更による可能性

本稿では、スペクトル包絡の生成で、初音ミクを基準としたスペクトル包絡の変形曲面を推定した。これは、スペクトルのどこをどのように変えれば初音ミク・アペンドが作れるのか、といった相対的な指標の推定に相当する。すなわち、この変形曲面をそのまま別の音源に適用できる**声色転写**の可能性を示唆している。実際、初音ミクから初音ミク・アペンドへ

の 6 種類の変形曲面を、そのまま鏡音リンに適用し、6 種類の「鏡音リン・アペンド」に相当する印象が得られたことを定性的に確認した。

しかし、このままでは汎用的な技術と限らない。鏡音リンで比較的うまくいくのは、同じ女性の声であること、それぞれ同じ歌声合成技術に基づいて作られた歌声であること、という二点において、両者が類似しているからである可能性が高い。実際、人間の男性の歌唱では適切な合成結果を得られなかった。これは、男女間の声道長の違いや個人性の違いがある場合に必要な、スペクトル包絡の非線形な対応付けを考慮していないことが原因である可能性が高い。したがって、声色転写の実現には、このような対応付けを考慮する必要がある。

また、初音ミクと鏡音リンが対応付いている保証もない。例えば、鏡音リンは初音ミク・アペンドの VIVID に対応付くものかもしれない。スペクトル変形曲面のような相対的な指標を活用するためには、そういった声色間の関係性も適切に推定できる必要がある。

5.2 応用 (2): 入力部（声色変化チューブの構成）の変更による可能性

本稿では、初音ミクと初音ミク・アペンドのような、歌唱者が同一の複数音源から声色変化を反映した歌声合成を行った。しかしここで、声色変化チューブを異なる歌唱者で構成することで、声質を動的に変化させて歌声合成できる可能性がある。また、本研究では既存歌声合成システムのパラメータ推定を行わなかったが、声色変化チューブを、例えば GEN パラメータを変えた複数の声から構成すれば、パラメータ推定に応用できる可能性がある。

6. おわりに

本稿では、これまで実現されていなかったユーザ歌唱からの声色変化の推定と、それを真似て歌声合成する VocaListener2 を提案した。本研究は、同一歌唱内における変動として、「声色変化」を活用するための新しい技術を示した点で意義があると考える。VocaListener2 によりユーザが手軽に人間らしい表現豊かな歌声を合成でき、さらには音高・音量・声色の多様な観点から、歌唱の表情付けが行えるようになった。

声質や声色は音高や音量と違い、物理量として単純に扱うことができず、未解決な課題も多い。そのような課題の一つとしては、適切な活用方法が明らかになっていないことが挙げられる。本研究では声色変化の活用について一つの具体例を示したが、今後は声色変化をモデル化して再利用する等、声色変化の新たな活用法について更なる検討をしていきたい。

本研究の根底には、文献 4) でも述べたように、「人間らしい歌唱」とは何かを解明し、より人間を知ることがある。本システムは、そうした歌声研究の基本ツールとしても貢献できる。例えば、VocaListener2 によって、音高や音量を真似た歌唱音声を様々な声色で用意できるようになったので、歌唱の個人性知覚に関する新しい知見が得られる可能性がある。

謝辞 本研究の一部は、科学技術振興機構 CrestMuse プロジェクトによる支援を受けた。また本研究では、RWC 研究用音楽データベース (音楽ジャンル RWC-MDB-G-2001) を使用した。

参考文献

[1] Cabinet Office, Government of Japan: Virtual Idol, *Highlighting JAPAN through images*, Vol. 2, No. 11, pp. 24–25 (2009). http://www.gov-online.go.jp/pdf/hlj_img/vol_0020et/24-25.pdf.

[2] 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).

[3] 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).

[4] 中野倫靖，後藤真孝：VocaListener: ユーザ歌唱を真似る歌声合成パラメータを自動推定するシステムの提案，情報処理学会研究報告 音楽情報科学 2008-MUS-75-9, Vol. 2008, No. 12, pp. 51–58 (2008).

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

[6] 剣持秀紀，大下隼人：歌声合成システム VOCALOID – 現状と課題，情報処理学会研究報告 音楽情報科学 2008-MUS-74-9，Vol. 2008, No. 12, pp. 51–58 (2008).

[7] 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， Vol. 48, No. 12, pp. 3637–3648 (2007).

[8] 森勢将雅：歌声を混ぜるインタフェース「e.morish」, http://www.crestmuse.jp/cmstraight/personal/e.morish/.

[9] Toda, T., Black, A. and Tokuda, K.: Voice conversion based on maximum likelihood estimation of spectral parameter trajectory, *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 15, No. 8, pp. 2222–2235 (2007).

[10] 大谷大和，戸田智基，猿渡 洋，鹿野清宏：STRAIGHT 混合励振源を用いた混合正規

分布モデルに基づく最ゆう声質変換法，電子情報通信学会論文誌， Vol. J91-D, No. 4, pp. 1082–1091 (2008).

[11] Schröder, M.: Emotional speech synthesis: A review, *Proc. Eurospeech 2001*, pp. 561–564 (2001).

[12] Iida, A., Campbell, N., Higuchi, F. and Yasumura, M.: A corpus-based speech synthesis system with emotion, *Speech Communication*, Vol. 40, Iss. 1–2, pp. 161–187 (2003).

[13] Tsuzuki, R., Zen, H., Tokuda, K., Kitamura, T., Bulut, M. and Narayanan, S. S.: Constructing emotional speech synthesizers with limited speech database, *Proc. ICSLP 2004*, pp. 1185–1188 (2004).

[14] 河津宏美，長島大介，大野澄雄：生成過程モデルに基づく感情表現における F_0 パターン制御規則の導出と合成音声による評価，電子情報通信学会論文誌， Vol. J89-D, No. 8, pp. 1811–1819 (2006).

[15] 森山　剛，森　真也，小沢慎治：韻律の部分空間を用いた感情音声合成，情報処理学会論文誌， Vol. 50, No. 3, pp. 1181–1191 (2009).

[16] Türk, O. and Schröder, M.: A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis, *Proc. Interspeech 2008*, pp. 2282–2285 (2008).

[17] Nose, T., Tachibana, M. and Kobayashi, T.: 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).

[18] Inanoglua, Z. and Young, S.: Data-driven emotion conversion in spoken English, *Speech Communication*, Vol. 51, Is. 3, pp. 268–283 (2009).

[19] 高橋　徹，西　雅史，入野俊夫，河原英紀：多重音声モーフィングに基く平均声合成の検討，日本音響学会研究発表会講演論文集（春季） 1-4-9，pp. 229–230 (2006).

[20] 川本真一，足立吉広，大谷大和，四倉達夫，森島繁生，中村　哲：来場者の声の特徴を反映する映像エンタテインメントシステムのための台詞音声生成システム，情報処理学会論文誌， Vol. 51, No. 2, pp. 250–264 (2010).

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

[22] Kawahara, H., Masuda-Katsuse, I. and de Cheveigne, A.: Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous frequency based on F0 extraction: Possible role of a repetitive structure in sounds, *Speech Communication*, Vol. 27, pp. 187–207 (1999).

[23] 西田昌史，有木康雄：音韻性を抑えた話者空間への射影による話者認識，電子情報通信学会論文誌， Vol. J85-D2, No. 4, pp. 554–562 (2002).

[24] 井上　徹，西田昌史，藤本雅清，有木康雄：部分空間と混合分布モデルを用いた声質変換，電子情報通信学会 技術研究報告 SP，Vol. 101, No. 86, pp. 1–6 (2001).

[25] Turk, G. and O'Brien, J. F.: Modelling with implicit surfaces that interpolate, *ACM Transactions on Graphics*, Vol. 21, No. 4, pp. 855–873 (2002).

[26] 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).