情報処理学会論文誌 Vol.54 No.6 1771–1783 (June 2013)

VocaListener2：ユーザ歌唱の音高・音量に加えて声色変化も真似る歌声合成システム

中野 倫靖[1,a)] 後藤 真孝[1,b)]

受付日 2012年1月25日, 採録日 2013年3月1日

概要：本論文では，ユーザの歌声からその声色（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．本システムは，我々が以前開発した音高と音量のみを真似て歌声合成する VocaListener の拡張であり，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，ユーザが歌唱において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．VocaListener2 を実現するために，まず VocaListener によってユーザ歌唱の音高，音量および音素（歌詞）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，構成された声色空間は聴取印象を反映しており，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．

キーワード：VocaListener，VocaListener2，歌声合成，声色変化

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

Tomoyasu Nakano[1,a)] Masataka Goto[1,b)]

Received: January 25, 2012, Accepted: March 1, 2013

Abstract: This paper presents a singing synthesis system, *VocaListener2*, that automatically synthesizes a singing voice by imitating timbre changes of a user's singing voice. The system is an extension of our previous *VocaListener* system which deals with only pitch (F_0) and dynamics (power). Most previous techniques for manipulating voice timbre have focused on voice conversion and voice morphing, and they cannot deal with intentional temporal timbre changes during singing. To develop VocaListener2, we first construct a *voice timbre space* on the basis of various singing voices that are synchronized under pitch, dynamics, and phoneme (lyrics) by using VocaListener. In this space, the timbre changes can be reflected in the synthesized singing voice. In our experiments with commercial singing synthesis systems, we found the constructed timbre space reflects the auditory impression and the timbre changes as well as the pitch and dynamics can be imitated.

Keywords: VocaListener, VocaListener2, singing synthesis, voice timbre changes

1. はじめに

本研究では，人間の歌い方を真似ることができる歌声合成システムの実現を目指す．この「真似る」アプローチは，複雑で時間のかかるパラメータ調整を手作業でしなくても，歌うだけで自然で表情豊かな歌声を手軽に合成できる点で有用である．自然な歌声を合成するためにパラメータ調整をする時間が減らせれば，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につながる．また，調整に関する知識を持たないユーザでも高品質な歌声合成結果を得ることができるようになり，様々な制作者にとって身近なツールが提供できれば，分野の発展に寄与できる．さらに，人間を真似るという高品質な歌声

[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)] t.nakano@aist.go.jp
[b)] m.goto@aist.go.jp

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る取り組みである．

2007年以降，日本では市販の歌声合成ソフトウェアが注目を集め，それを用いて楽曲制作を行う一般ユーザや音楽家（プロ）が増えた．そうした楽曲が収録された音楽CDが市販され，商業音楽ヒットチャート上位にランクインする等，楽しむリスナも増加している [1]．また，そうして創られた作品の多くは，動画コミュニケーションサービス「ニコニコ動画*1」に投稿されたことで，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を介した大規模な協調的創造活動につながってきた [2], [3]．また，さらに，CG映像によるライブが日本国内や海外*2で開催される等，音楽の鑑賞における新たな形態も登場し始めている現状がある．しかし，人間らしい自然な歌声を合成しようとすると難易度が高く，適切な知識や時間をかけた調整が必要だったために，誰でも容易に使いこなせるものではなかった．

そこで我々は以前，入力としてユーザが歌声を与え，その音高と音量を真似るように歌声合成できるシステムVocaListenerを開発した [4], [5]．しかし，これまでのVocaListener（本論文では以下，VocaListener1と呼ぶ）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．本論文ではそれを拡張し，ユーザ歌唱の声色変化も歌声合成結果に反映できるVocaListener2を提案する．歌唱中の声色変化も真似て，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，歌声や話声の声質変換や声質モーフィングに関しては様々な研究がなされてきた（2.2節で後述）が，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（せいしつ，voice quality）」という用語 [6], [7], [8] 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（評語）の違い等，多様な意味合いで使われている．そこで，本研究ではそういった発声様式や聴感上の印象における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（こわいろへんか，voice timbre changes）」という単語を用いる．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 [9] があった．Vocaloidでは，スペクトル情報を制御する複数の数値パラメータを各時刻で調整することで，歌声のスペクトルを操作して声色変化をともな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ごとに一括で変更したり，大まかに変更したりしていた．

そのような問題を解決するために，本研究では前述の「真似る」アプローチによる声色変化制御を実現する．そのために，まず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，それらの歌声すべ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ることができないため，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章で用語の定義（声質および声色変化）と関連研究について述べ，3章で我々が以前開発したVocaListener1の問題点を提起する．続いて，4章で問題点を解決するVocaListener2の実現方法について述べた後，5章で実験を行う．最後に，6章で本研究の意義と今後の課題について述べる．

2. 用語の定義と関連研究

従来，声質に関する様々な研究がなされてきた．本章では，声質と声色変化という用語を定義したうえで，声質変換やモーフィングに関する従来研究について説明し，本研究の位置付けを示す．

2.1 声質と声色変化

声質とは，「音声波から知覚される韻質（いんしつ，phonemic quality）以外の聴覚上の特質」と定義される [8]．狭義には発声の仕方（喉頭制御）の違いに起因する聞こえの違い（緊張した声，ささやき声等）を表し，広義には音声器官の生理的・物理的違いによる話者の個人差に起因する聞こえの違いをも表す [7], [8]．ここで，韻質とは音声に含まれる情報のうち言語情報にかかわるものであり，韻質以外の情報を声質と呼ぶ．アクセント等の超分節音素も言語情報を担っているが，超分節音素が担う情報を韻質から区別して韻律（いんりつ，prosody）と呼ぶこともあり，その場合，韻質は分節音素が担う言語情報だけを指すことになる [8]．一方，声質を個人性に限定して「話しているとき，継続的に存在する特質*3」と定義されることもある [6]．声質の包括的な説明や知見としては，文献 [6], [7], [10] が参考になる．

これに対して本研究では，「声色」という用語を，「同一人物の音声波（話声および歌声）から知覚される韻質（音韻

*1 http://www.nicovideo.jp/
*2 アメリカとシンガポールにて開催（2012年1月時点）．
*3 The term ‘voice quality’ refers to those characteristics which are present more or less all the time that a person is talking: it is a quasi-permanent quality running through all the sound that issues from his mouth.（文献 [6] のp.91より抜粋）

および韻律）以外の特質のうち，発声の仕方の違いに起因する聴覚上の特質（きこえ）の違い」という意味で用い，それが時間変化するという意味で「声色変化」を用いた．声色という単語は，Laver の声質に関する定義（広義）[7] [*4] を参考にして決定した[*5]．

2.2 関連研究と本研究の位置付け

歌声を対象として声質・声色変化を扱った研究としては，歌声合成における声質制御がある．これは，市販のソフトウェアにおけるユーザによる手作業（マウス）での数値パラメータ調整 [9] が一般的で，歌声合成システムを声質変換技術によって拡張する研究 [13] もある．また同一の個人性で複数の声色の音源（歌声合成用の歌声データベース）も市販されており，たとえば，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）[*6]」は，「初音ミク[*7]」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．

そのほか，ラップ歌唱の声質変換 [14]，混合音中の歌声（音楽 CD 等の背景音楽付きの歌唱）に対する声質変換 [15] や，声道断面積関数に基づく声質変換 [16] 等が研究されている．また 2 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 [17], [18]，感情を変えて歌った同一歌唱者の複数の歌唱を混ぜる感情モーフィング [19] がある．

一方，話声を対象として声質・声色変化を扱った研究には，声質変換を対象として様々な研究があり，主に，異なる話者の話声間の変換モデルを構築する統計的変換手法がさかんに研究されている．具体的には，コードブックマッピング法 [20] をはじめ，ニューラルネットワークを用いる方法 [21], [22] や，隠れマルコフモデル（HMM）のパラメータ適応に基づく方法 [23], [24]，混合正規分布モデル（GMM）に基づく変換法 [25], [26], [27], [28], [29], [30] がある．その中でも GMM に基づく変換法は，その柔軟性から近年多くの研究があり，話者性の変換だけでなく，食道音声 [31] や人工喉頭音声 [32] を入力とする声質変換の研究にも応用されている．ここで，変換モデルの構築の際に発話内容が同一の入出力音声対（パラレルデータ）を必要としない手法も提案されている [33], [34]．また，多数の話者の話声データを活用するための固有声変換法に基づく様々な研究もあり [35], [36], [37], [38]，異なる話声の変換に関する効果的な手法が提案されてきた．

[*4] Voice quality is conceived here in a broad sense, as the characteristic auditory colouring of an individual speaker's voice, and not in the more narrow sense of the quality deriving solely from laryngeal activity. （文献 [7] の p.1 より抜粋）

[*5] 声質の意味で，「声音（こわね）[8]」や「声の音色（voice timbre）[11], [12]」という用語が使われることもあった．

[*6] http://www.crypton.co.jp/cv01a/

[*7] http://www.crypton.co.jp/mp/pages/prod/vocaloid/cv01.jsp

そのほか，感情音声合成に関する研究があり [39], [40], [41], [42], [43], [44], [45], [46], [47]，話声の韻律や話速を扱うものが多いが，感情変化にともなう声質変換 [39], [40], [41], [42], [43] 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 [48] や，母音をモーフィングすることで話者性を変換する研究 [49], [50], [51]，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研究 [52] もある．

しかし，いずれの研究も，歌唱中の声色変化を制御することはできないか，可能性はあっても運用困難であった．前述したように，歌声合成ソフトウェアにおける声質パラメータは，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い．また，複数の音源をフレーズごとや曲ごと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る．声質変換やモーフィングに関しては，その多くが異なる複数の音声間での変換を対象としており，そのままでは歌唱中の変化を表現できない．ここで，個人の複数の声色からモデル学習することで，声質変換やモーフィングを応用して歌唱中の声色変化を制御できる可能性がある．ただし，その場合でも，前に述べた手作業によるパラメータ制御の同様に，いつどういった声色で歌わせるか，といった声色変化を指定する制御の実現が困難である．

それに対して本研究では，「このように歌ってほしい」という方針を歌って表現できる．声色変化は，物理量として単純に扱うことができないため，歌を真似るアプローチは，表情豊かな歌い方を直感的に合成できる可能性がある．

3. VocaListener1 の機能とその問題点

本章では VocaListener1 の概説と，VocaListener2 の実現課題について述べる．

3.1 VocaListener1：ユーザ歌唱の音高と音量を真似る歌声合成システム

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 [53] に比べて向上し，歌声合成システムやその音源（歌声合成用の歌声データベース）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，音符ごとに割り当てる作業はほぼ自動で行える．音符の割当てでは，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．システム構成図や処理内容の詳細や，性能の具体的な数値に関しては文

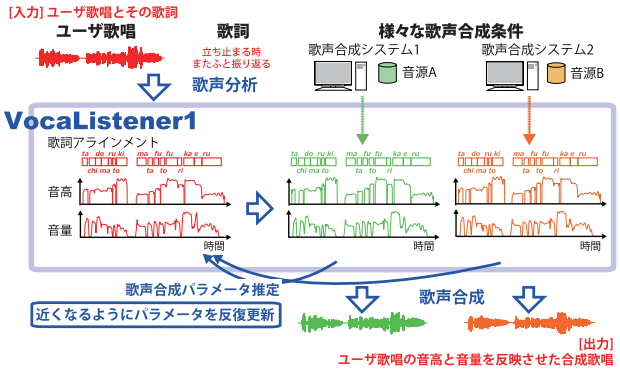

図 1 VocaListener1 によるユーザの歌声とその歌詞を入力とした，音高と音量に関する歌声合成パラメータ推定の概要

Fig. 1 Overview of VocaListener1, which iteratively estimates parameters of pitch and dynamics for singing synthesis from the user's singing voice and the song lyrics.

献 [4] 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[*8]や動画コミュニケーションサービス『ニコニコ動画』[*9]上で視聴できる．

3.2 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 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，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る可能性が高く，仮に同じ名称のパラメータでも，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採用できない．たとえば，ヤマハ株式会社の Vocaloid と Vocaloid2 [1] では，声質に関する操作可能なパラメータの名称とその処理内容が一部異なる．

4. 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 VocaListener1 の問題点をふまえ，VocaListener2 の実現課題と解決方策を説明する．

4.1 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 2 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，本論文では，スペクトル包絡の形状の違いとして定義する．**図 2** に，Vocaloid2 における，同一の個人性を持つ複数の音源（音源 A）で合成した音と，

[*8] http://staff.aist.go.jp/t.nakano/VocaListener/index-j.html
[*9] http://www.nicovideo.jp/mylist/7012071/

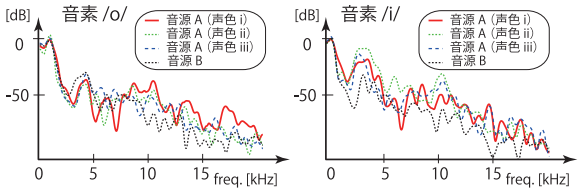

図 2 音素の違いと歌唱者の違いによるスペクトル包絡形状の違いの具体例

Fig. 2 Examples of differences in the spectral envelope due to phonetic and individual differences.

それとは異なる個人性を持つ音源（音源 B）で合成した音から推定したスペクトル包絡の例を示す．

スペクトル包絡形状の違いには，図 2 に示すように，音素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歌唱表現としての声色変化といえる．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．

4.2 解決方針

前節で述べた実現課題 (1) を解決するために，本論文では市販されている歌声合成ソフトウェアにおいて，同一の個人性を持つ複数の声色の音源を用いて声色変化を表現する．具体的には，それら複数の音源から 1 度歌声合成を行い，その結果が滑らかに切り替わって変化するスペクトル包絡を生成する．本論文ではこれ以降，このような複数の音源（合成対象）を指して「声色セット」と呼ぶ．

続いて，実現課題 (2) を解決するために，まず，ユーザ歌唱と声色セットのスペクトル包絡を「声色空間」へ射影する．ここで声色空間とは，声色の違いが表現された空間であり，すべての歌唱が各時刻において，それぞれ声色空間上の 1 点に対応付けることができるものであるとする．そのような射影によって，ユーザ歌唱は声色空間上の時間変化する軌跡として表現できる．最後に，その軌跡から，声色セットに基づく新たなスペクトル包絡を生成することができれば，ユーザ歌唱の声色変化を真似て歌声合成できる．

本節では以降，具体的な声色空間の定義と扱いについて述べ，次節以降で詳細な実現方法を述べる．空間上の 1 点からのスペクトル包絡の生成については，4.7 節で述べる．

4.2.1 本論文における声色空間の定義

理想的な声色空間は，スペクトル包絡から声色の違い以外（主に，個人性と音素の違い）を抑制して得られる空間である．その空間では，異なる歌唱者による 2 つの歌声があるとき，それらの声色に関する印象が同じであるならば，声色空間内で同じ点に射影される．このような理想的な声色空間は，多様な個人性を持つ多数の歌声と，そのそれぞれについて声色が多様に含まれていれば，構成できる可能性がある．しかし，そのような多種多様な大量の歌声デー

タを用意することは難しい．

そこで本論文では，個人性と声色の違い以外（主に，音素の違い）を抑制するように声色空間を構成する．ただし，ある音源とその別の声色の音源との位置関係が，ユーザ歌唱とその別の声色との位置関係と同じ空間であるとする．いいかえると，ある音源のある時刻のスペクトル包絡をベクトル $\mathbf{z}$，ユーザ歌唱のスペクトル包絡をベクトル $\mathbf{u}$ として，ある声色 t_1 と別の声色 t_2 について，$\mathbf{z}_{t_1} - \mathbf{z}_{t_2} = a(\mathbf{u}_{t_1} - \mathbf{u}_{t_2})$ であると仮定する．ここで a はスケーリング係数を表す．そのうえで，声色空間上での声色セット中の基準となる声色とユーザ歌唱における基準となる声色の位置関係を正規化して合わせること（シフトとスケーリング）ができれば，声色空間上でのユーザ歌唱の軌跡に対応する声色の変化が推定できる．

このような空間は，声色セットと複数人による歌声を用意すれば構成できる可能性がある．次項では複数人の歌声の必要性について考察し，続いて声色空間の構成方法として，データの用意方法と音素の違いの抑制方法を述べる．

4.2.2 声色空間の構成に関する考察

ある音源とその別の声色の音源との位置関係が，ユーザ歌唱とその別の声色との位置関係となるべく同じ空間を構成するためには，表現能力の高い声色空間を構成することが必要である．したがって，空間を構築する音源は多様なものであると望ましい．

なぜなら，多様な複数の歌声から声色空間を構成すれば，スペクトル包絡の形状が類似した歌声（似ている声）は近くの領域に射影され，個人性の違いは射影される領域の違いとして表現されると考えられる．今回は，声色セットの各声色の違いを，個人性の違いととらえていることに相当しており，それらの相対的な位置関係は，空間を構成する歌声に影響を受ける．つまり，声色の違いを個人性の違いを使って説明しようとしているといえる．したがって，相対的な位置関係が適切になる声色空間を構成するためには，多様な個人性が含まれている方が良いと仮定する．

逆に，もし単一の声色セットのみで声色空間を構成してしまうと，男性のユーザが女性の声色セットを用いる場合等には，適切な位置関係が得られない可能性があると考えられる．

4.2.3 声色空間の構成方法

個人性と声色の違い以外（主に，音素の違い）を抑制するように声色空間を構成するために，まず VocaListener1 を用いて，複数の歌唱者の音源 I 個からユーザ歌唱を真似て，時刻が同期した歌声を自動的に合成する．ここでは，合成対象となる声色セットも J 個含まれるとする（$J \leq I$）．これによって，各時刻において音高・音量・音素が同期した歌唱が得られるため，ユーザ歌唱および I 個の合成結果を合わせた $K(=I+1)$ 個の歌声を活用して，個人性や声色の違い以外の成分を抑制した声色空間を構成する．このようにすることで，曲ごとの声色空間が効率的に構成できる（4.5 節で詳しく説明）．続いて，声色セットによる合成結果（同期した歌唱）の，声色空間上における J 個の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．ただし，声色変化チューブは，説明の都合上の概念であり，実際に何らかのパラメータによって構築するものではない．

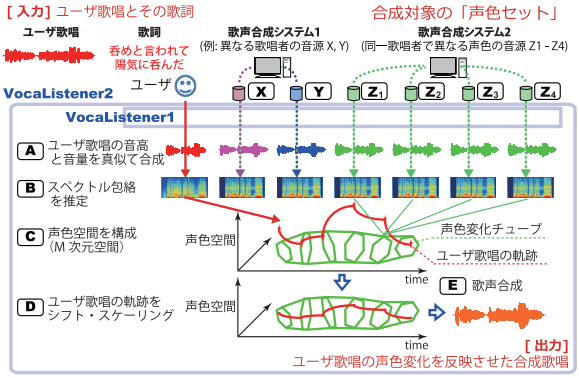

図 3 VocaListener2 における処理の流れ．ユーザ歌唱の音高・音量・声色変化を真似るように歌声合成する

Fig. 3 Overview of VocaListener2, which automatically synthesizes a singing voice by imitating pitch, dynamics, and timbre changes of a user's singing voice.

声色空間を M 次元空間とすると，声色セットとして，各時刻 t において J 個の M 次元ベクトル $\mathbf{z}_j(t)$ $(j = 1, 2, ..., J)$ がその空間上に存在する．声色空間は，多様な歌声を合成して活用することで，個人性と声色セットにおける声色の違い以外の成分を抑制したものであり，それら J 個の点 $\mathbf{z}_j(t)$ に囲まれた内側やその付近では，声色セットの個人性を持っている可能性が高い．逆にその外側や離れた位置では，異なる個人性が表現されている可能性がある．そこで，その内側を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．

これ以降，特に明記しない限り，処理は離散的なものとして扱い，その時間単位は 1 ms とする．

4.3 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 は合成対象となる声色セットであり，この例では $J = 4$ である．まず，ユーザ歌唱から VocaListener1 を

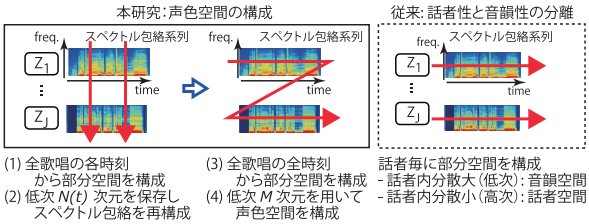

図 4 部分空間に基づいた声色空間の構築と従来研究 [55], [56] との違い

Fig. 4 Difference between our method for voice timbre space construction by a subspace method and previous method [55], [56].

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声が得られる．

続いて，それぞれの歌声を分析し，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，4.1 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等による 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実現方法を説明する．具体的な分析条件は次章で述べる．

4.4 歌声分析：歌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 [54] 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 [17]．

4.5 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・音素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．ただし実際には，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．したがって，時刻ごとに主成分分析を行った結果，時刻ごと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，声色変化と個人性の寄与が大きな空間として考えることができる．

部分空間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 [55] や声質変換 [56] において有効性が確認されている．従来研究 [55], [56] では，話者ごと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，本研究ではそれを時刻ごとに行う．しかしそのままでは，各時刻で異なる空間が構成されることになり，全時刻を統一的に扱えない．そこで，まずは時刻ごとの部分空間における低次 $N(t)$ 次元のみを保存して，元の空間に戻すことで，個人性と声色変化に寄与する成分以外を抑制する．続いて，全歌唱の全時刻を用いて 1 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．ここで，全歌唱の全時刻に関する主成分分析のみを行うと，変動が大きな低次部分空間には音素の違いが多く残ることが考えられるため，このような処理を行った．

処理の具体内容について説明する．まずは，ある時刻 t についてのみ考える．歌声 k の周波数 f のスペクトル包絡を $S_{t,k}(f)$ とする．ただし現在の実装では，それぞれの t と k におけるスペクトル包絡に対して，離散コサイン変換を行い，その低次成分のみを用いる次元圧縮を行った後で以降の処理を行う．これらのベクトルから主成分分析を行って射影した結果，各主成分（次元）d における値が $C_{t,k}(d)$ として得られる．ここで，事前に決めた累積寄与率の下限 R_e を超える次元数を $N(t)$ とすると，それに基づいて，次式のように高次主成分を 0 とする．

$$\hat{C}_{t,k}(d) = \begin{cases} C_{t,k}(d) & (d \le N(t)) \\ 0 & (d > N(t)) \end{cases} \tag{1}$$

この $\hat{C}_{t,k}(d)$ をスペクトル包絡に逆射影して，スペクトル包絡 $\hat{S}_{t,k}(f)$ を得る．以上のような主成分分析を時刻 t ごとに行う．続いて，このようにして得られた歌声 k の時刻 t，周波数 f のスペクトル包絡 $\hat{S}_{t,k}(f)$ から，すべての t と k のベクトルに対して主成分分析を行って，低次 M 次元の空間を声色空間として得る．

以上のような処理によって，全歌唱者の全時刻におけるスペクトル包絡が同じ空間上で扱えるだけでなく，音素の

違い等の文脈にとも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．さらに，このような個人性と声色の違い以外の成分を抑制する処理は，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．

4.6 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，本論文では単純な方法として，ユーザ歌唱の軌跡と声色セットの軌跡すべてのそれぞれで，次元 d ごとに 0〜1 の値となるように正規化することで対応付けた．これはシフト・スケーリング操作で位置合わせしていることに相当する．具体的には，時刻 t，次元 d における声色空間上でのユーザ歌唱の軌跡を $u_d(t)$，また各声色 j の軌跡を $z_d(j,t)$ として，以下のように実現した．

$$\hat{u_d}(t) = \frac{u'_d(t) - \min u'_d(t)}{\max u'_d(t) - \min u'_d(t)} \tag{2}$$

$$u'_d(t) = u_d(t) * h(t) \tag{3}$$

$$\hat{z_d}(j,t) = \frac{z_d(j,t) - \min z_d(j,t)}{\max z_d(j,t) - \min z_d(j,t)} \tag{4}$$

ここで，$\min z_d(j,t)$ と $\max z_d(j,t)$ は，全時刻全声色の中での最小と最大を意味する．また，急峻な変化を抑制するために，時刻 t，次元 d におけるユーザ歌唱の軌跡 $u_d(t)$ については平滑化を行った．現在の実装では，平滑化フィルタ $h(t)$ として，時間方向に 200 ms（$N = 200$）の長さを持つ三角フィルタを畳み込んだ（$*$ が畳み込みを表す）．

ここで本手法では，4.2.1 項で述べたように，入力の歌唱の声色変化は，声色空間における声色セットの各音源の位置関係と相対的に同じと仮定している本論文では，ユーザ歌唱とその別の声色との位置関係とがなるべく同じとなる声色空間を構成するために，市販されているすべての音源を使うことで対応した．しかし，入力された歌声と声色空間を構成する音源の種類等によっては，必ずしも適切に動作しない可能性もある．今後，入力歌唱と声色セットの音源の相対的な位置関係の考慮や，回転や何らかの非線形変換等も導入することで，より柔軟な声色変化の反映方法へ発展させることができる可能性がある．

4.7 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 1 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，ある時刻におけるある声色セットの配置を**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

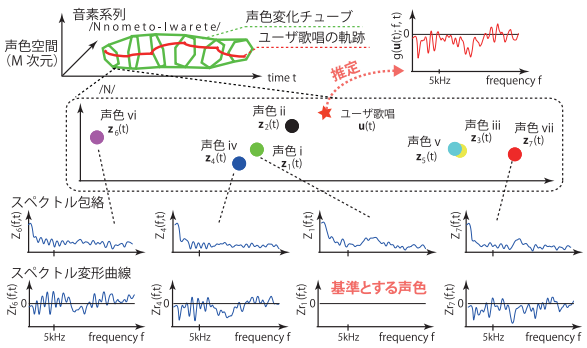

図 5 声色空間の推定結果例．各点には，ユーザ歌唱と合成対象（声色セット）の各音源のスペクトル包絡が対応付いている

Fig. 5 Example of an estimated spectral envelope from vectors for user's singing and target timbre voices in the voice timbre space.

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，2 つ以上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 [17], [48]．しかしここでは，歌唱者・音高・音素が同一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ごとの強調・抑制処理のみで声色変換が可能であると仮定する．

ここでは，スペクトル包絡をそのまま使うのではなく，標準的な声を基準としてそこからの変形比率として表し，この比率をまず時刻ごとに推定する．これによって，急峻すぎる変化や望まない変形等を抑制したり，逆に強調したりする処理を導入しやすくなるため，品質を保ったまま変形できる．本論文では，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 [57] を応用して適用する．

まず，時刻 t，周波数 f における各声色のスペクトル包絡を $Z_j(f,t)$（$j = 1, 2, ..., J$），その $Z_1(f,t)$（$j = 1$，基準の声色）に対するスペクトル変形曲面を $Zr_j(f,t)$ として次のように定義する．

$$Zr_j(f,t) = \log\left(\frac{Z_j(f,t)}{Z_1(f,t)}\right) \tag{5}$$

ここでは対数をとり，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とることを許容する．

入力として，声色空間における M（$= 3$）次元のある座標 $\mathbf{x}(t)$ が与えられたときに，声色空間上での各声色の軌跡を $\mathbf{z}_j(t)$ とすると，スペクトル変形曲線を計算する関数（モデル）$g(\mathbf{x}(t); f, t)$ は

$$g(\mathbf{x}(t);f,t)=\sum_{j=1}^{J}\left(w_{k,f,t}\cdot\phi\left(\mathbf{x}(t)-\mathbf{z}_j(t)\right)\right)+a_{0,f,t}+\sum_{m=1}^{M}a_{m,f,t}\cdot x^{(m)} \tag{6}$$

となる．ここで $w_{k,f,t}$，$a_{0,f,t}$ および $a_{m,f,t}$ が未知のモデルパラメータ，$\phi(\cdot)$ は，ベクトル間の距離を表す関数であり，本論文では $\phi(\cdot)=|\cdot|$ を用いる．これらの未知パラメータを推定するために，各声色の軌跡とそのスペクトル包絡は，すでに正解が与えられているため，

$$g(\mathbf{z}_j(t);f,t)\ =\ Zr_j(f,t) \tag{7}$$

といった制約条件を与えることができる．これは，指定した座標（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ことを意味する．このような制約条件を満たしながら，式 (6) によって推定された関数 $g(\mathbf{x}(t);f,t)$ が滑らかになるように（二次導関数の二乗の和が最小になるように）モデルパラメータを推定するために，以下のような線形システムとして表して解く（詳細は文献 [57] を参照）．

$$\begin{bmatrix}\phi_{11}\cdots\phi_{1J} & 1 & z_1^{(1)} & z_1^{(2)} & z_1^{(3)}\\ \phi_{21}\cdots\phi_{2J} & 1 & z_2^{(1)} & z_2^{(2)} & z_2^{(3)}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \phi_{J1}\cdots\phi_{JJ} & 1 & z_J^{(1)} & z_J^{(2)} & z_J^{(3)}\\ 1\cdots 1 & 0 & 0 & 0 & 0\\ z_1^{(1)}\cdots z_J^{(1)} & 0 & 0 & 0 & 0\\ z_1^{(2)}\cdots z_J^{(2)} & 0 & 0 & 0 & 0\\ z_1^{(3)}\cdots z_J^{(3)} & 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}w_1\\ w_2\\ \vdots\\ w_J\\ a_0\\ a_1\\ a_2\\ a_3\end{bmatrix}=\begin{bmatrix}Zr_1\\ Zr_2\\ \vdots\\ Zr_J\\ 0\\ 0\\ 0\\ 0\end{bmatrix} \tag{8}$$

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 $w_{k,f,t}$ と $a_{0,f,t}$，$a_{m,f,t}$ を用いて，声色空間上でのユーザ歌唱 $\mathbf{u}(t)$ から，その座標に対応するスペクトル変形曲面 $g(\mathbf{u}(t);f,t)$ を次式で得ることができる．

$$g(\mathbf{u}(t);f,t)=\sum_{j=1}^{J}\left(w_{k,f,t}\cdot\phi\left(\mathbf{u}(t)-\mathbf{z}_j(t)\right)\right)+a_{0,f,t}+\sum_{m=1}^{M}a_{m,f,t}\cdot x^{(m)} \tag{9}$$

ここで，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．現在の実装では，2 次元の平滑化フィルタ $H(f,t)$ として，時間方向に 100 ms，周波数方向に約 100 Hz の長さを持つ二次元三角フィルタを畳み込む．最後に，基準の声色のスペクトル包絡 $Z_1(f,t)$ にこのスペクトル変形曲面を用いて変形し，合成用スペクトル包絡 $Z_u(f,t)$ を得る．ここで場合によっては，合成の不自然さを減らすために，高域の値は重みを付けて反映を抑制し，声色変化チューブ外にユーザ歌唱が存在した場合等の影響を低減させる．具体的には，通過域カットオフ周波数 F_L と阻止域カットオフ周波数 F_H を定めて，合成用スペクトル包絡 $Z_u(f,t)$ を以下のように求める．

$$Z_u(f,t)\ =\ Z_1(f,t)\times\hat{R}_z(f,t) \tag{10}$$

$$\hat{R}_z(f,t)\ =\ \beta\cdot(R_z(f,t)*H(f,t))+(1-\beta)\cdot R_z(f,t) \tag{11}$$

$$R_z(f,t)\ =\ (1-R_g(f,t))\cdot\exp\left(g(\mathbf{u}(t);f,t)\right)+R_g(f,t)\cdot 1 \tag{12}$$

$$R_g(f,t)\ =\ \begin{cases}0 & (f<F_L)\\ \frac{\alpha}{F_H-F_L}\cdot(f-F_L) & (F_L\le f\le F_H)\\ \alpha & (F_H<f)\end{cases} \tag{13}$$

ここで，それぞれの値は実験的に，$F_L=5000$，$F_H=6000$，$\alpha=0.3$，$\beta=0.95$ とし，$*$ は畳み込みを表す．合成用スペクトル包絡 $Z_u(f,t)$ を STRAIGHT [54] 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．ここで，STRAIGHT の合成で必要となる非周期性指標については，基準となる歌声（$j=1$）の分析結果をそのまま用いた．

4.8 インタフェース：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，表現の幅を広げるため，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 3 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 2 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．

5. 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．

表 1 各時刻における 18 種類の歌声のスペクトル包絡を主成分分析した際の，累積寄与率 R_e%を超える次元数 $N(t)$ の平均と標準偏差

Table 1 Means and standard deviations of $N(t)$-dimension for the cumulative contribution ratio to become more than R_e% of principal components analysis (PCA) for each frame of 18 singing voices.

累積寄与率 R_e [%]	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

5.1 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 [58] No.91「大漁船」を用いて行った．これ以降，歌声はサンプリング周波数 44.1 kHz のモノラル信号を扱い，処理の時間単位は 1 ms 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 [1] を採用し，その応用商品として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すべてにデフォルト値を用いた）．これらのソフトウェアには，それぞれ個人性を識別するための名前が設定されているため，以下に各名称をかぎ括弧「」で囲んで列挙する．用いた男性歌唱が 3 種類であり，「KAITO」（Vocaloid1），「がくっぽいど」，「氷山キヨテル」（以上，Vocaloid2）を用いた．また，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 種類であり，「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）を用いた．以上をまとめると，声色空間の構成には上記の 17（$= K$）個による歌声とユーザ歌唱を合わせた 18（$= L = K + 1$）の歌声を用いる．このように声色空間の構成に複数の歌声を用いるのは，4.2 節で述べたように，表現能力の高い声色空間を構成するためである．また，「初音ミク」と「初音ミク・アペンド」の 7 種類（$= J$）を声色セットとして合成対象の歌声として扱う．

またこれ以降，初音ミクと初音ミク・アペンドを区別するために，初音ミクを NORMAL，初音ミク・アペンドをそれぞれ DARK，LIGHT，SOFT，SOLID，SWEET，VIVID として示す．

それぞれの歌声を STRAIGHT によって，各時刻でスペクトル包絡を推定する際には，分析窓長は 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 スペクトルをよく再現できたためである．

その後，時刻ごと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 R_e が 80%を超える次元数を用い（次元数 $N(t)$ は時刻ごとに可変），全歌唱の全時刻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（この場合の累積寄与率を R_a とする）．またこれらの処理は，すべての歌唱で F_0 が存在する有声区間のみを用いた．

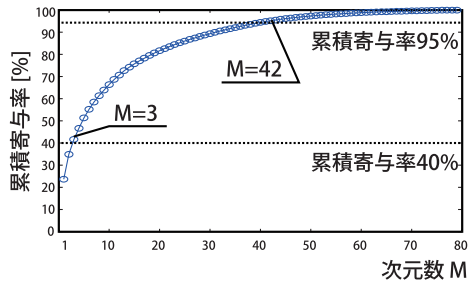

図 6 18 種類の合成歌唱のスペクトル包絡をすべて用いて主成分分析した際の，累積寄与率 R_a%を超える次元数 M の関係

Fig. 6 Relation between the dimension M and the cumulative contribution ratio R_a of PCA for all frames of 18 singing voices.

5.2 実験 A：提案手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 秒）を用いて，部分空間に基づいた声色空間の構成に関してその特性を確認する．

まず，各時刻における 18 種類の歌声のスペクトル包絡を主成分分析した場合の，累積寄与率 R_e と次元数 $N(t)$ の対応関係を調べる．その $N(t)$ の全時刻での平均と標準偏差を求めた結果を**表 1** に示す．次に，時刻ごとの部分空間における累積寄与率 R_e が 80%を超える低次 $N(t)$ 次元のみを保存して，元の空間に戻した後，18 種類の歌声のスペクトル包絡の全時刻を主成分分析する．その場合の，累積寄与率 R_a と次元数 M の関係を**図 6** に示す．ここで，累積寄与率 R_e が 80%を超える次元数 $N(t)$ の平均は 4.18 次元であり，全歌唱の全時刻における主成分分析では，M（$= 3$）次元を用い，その累積寄与率は $R_a = 42.3$%であった．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時刻における平均ベクトルとその分布を**図 7** に示し，**図 8** に，構成された M（$= 3$）次元の声色

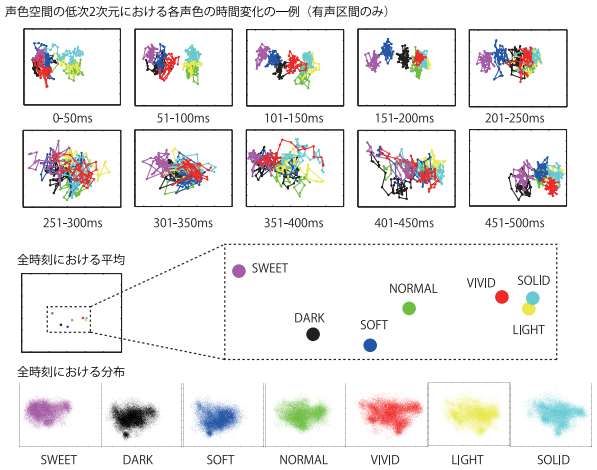

図 7 構成された声色空間の低次 2 次元の例：各声色の時間変化（上段）および分布（中段），声色ごとの全時刻における平均（下段）

Fig. 7 An example of the first two principle components of constructed voice timbre space: fluctuations (top) and distributions (middle) of each voice timbre, and their averages of all frames (bottom).

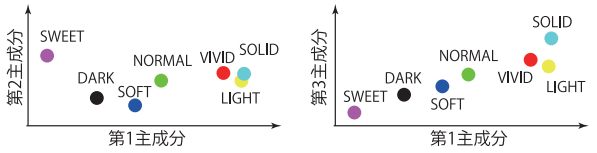

図 8 構築された声色空間の低次 3 次元における声色ごとの全時刻における平均ベクトル

Fig. 8 Constructed 3-dimensional voice timbre space and average vectors of the first three principle components for each voice timbre.

空間とその平均ベクトルを示す．

表 1 からは，スペクトル包絡の形状の多く（累積寄与率 $R_e = 95\%$）が平均 7.7 次元で表現できるといえる．ただし，全歌唱の全時刻に着目すると，歌詞（音素）の違いを含むことから，スペクトル包絡形状の分散は大きくなる（図 6）．しかし，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．実際，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．上位 3 次元に着目しても，図 8 の結果から，声色空間上で各声色の平均ベクトルは比較的分離している．また，定性的ではあるが，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を反映するような配置となっていた．具体的には，VIVID と SOLID と LIGHT がそれぞれ類似しており，DARK と SOFT も類似していた．また，SWEET はそれらとは大きく異なっていた．構成された声色空間とそれに対応する合成結果はホームページ http://staff.aist.go.jp/t.nakano/VocaListener2/index-j.html から確認できる．以上の結果から，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効率的に表現できるといえる．

5.3 実験 B：ユーザ歌唱の声色変化の推定結果の確認

本実験では，ユーザ歌唱の声色変化の推定結果を評価する．そのために本論文では声色セットから NORMAL を除いた 6 種類（DARK，LIGHT，SOFT，SOLID，SWEET，VIVID）を入力として与え，その声色変化を推定して結果を考察する．このように，声色空間の構成に用いられる声色セットと同様の声色での歌声合成結果を入力として用いることで，推定結果を一部評価できる．

本実験の実施は，同一人物の声で，かつ聴覚上のきこえの違いとしての声色変化がある歌唱を入力として与える必要がある．ここで，実験で用いる声色セットに関しては，「各音源が制作される際に，人間の聴覚上のきこえの違いが考慮されている」という仮定に基づいている．さらに，これらの音源を用いて合成された歌声は，インターネット上の動画コミュニケーションサイト等で多数のユーザが実際に使用している結果を聴取できる．それと，著者自身が実際に合成した際の聴取印象からも合わせて，同一人物の声でありながら聴覚上のきこえの違いとしての声色変化があると判断した．

まず便宜上のユーザ歌唱として「大漁船」の男性歌唱（55 秒）を VocaListener1 [4], [5] で音高と音量，歌詞を真似て合成した 6 種類の歌唱について，それを手作業で切り貼りした歌声を対象とする．このように，VocaListener1 を用いることで，音素のタイミングおよび音高と音量の変化が同一で声色のみが異なる歌声が生成でき，切り貼りによって，擬似的にフレーズごとに声色が変化する歌声を生成できる．VocaListener1 を用いて男性歌唱（人間）から生成したのは，考察を適切に行うために，ユーザ歌唱以外の条件（音高や音量変化）を実験 A と近付けることで，声色空間の特性をなるべく同一にすることが目的である．

ただし，声色空間を構成する際にもそれら 6 種類の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合成した歌唱も用いる．これによって，声色空間における入力の歌唱の声色変化と声色セットにおいて，相対的な位置関係が同じであるという仮定（4.2.1 項，4.6 節）を確認する．具体的には，声質に関するパラメータ（GEN）をデフォルト値（= 64）として切り貼りした歌唱（Closed 実験）とそれを 90 に変更して 2 半音下げた歌唱（Open 実験）の 2 種類を入力した．ここで，GEN パラメータは Vocaloid2 におけるスペクトル形状の変形を行うものであり，GEN が変更されて合成された音は，声道長が変更された声となるような聴取印象を受けた．

図 9 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，提案した声

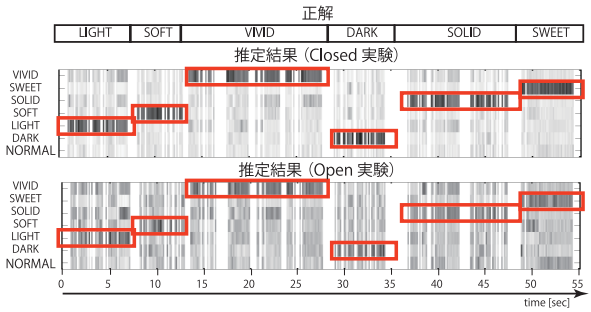

図 9 実験 B における Closed/Open 実験の結果．推定結果と声色空間上での各声色とのユークリッド距離（濃いほど距離が近い）

Fig. 9 Results of the closed/open experiment in experiment B. The Euclidean distance between each voice timbre and the estimated trajectory in the voice timbre space is shown (the nearer to the timbre, the darker in the color).

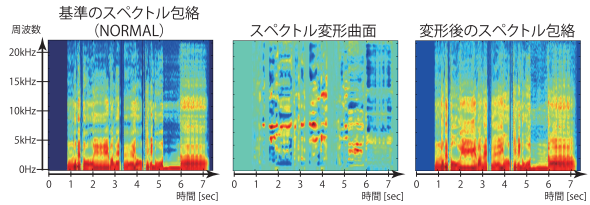

図 10 推定されたスペクトル変形曲面と，基準とする声色（NORMAL）にそれを適用したスペクトル包絡の例

Fig. 10 An example of an estimated spectral transform surface and the result of applying it to the standard spectral envelope (NORMAL).

色空間の構成とシフト・スケーリング操作によって，正解がほぼ推定できたことが分かる．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 が影響を受けていたりすることがあった．図 7 の結果を見ると，これらの声色が近くに配置されていることから，そもそもこれらのスペクトル包絡形状が類似していることに原因があると考えられる．ここで，スペクトル包絡（スペクトル変形曲面）の推定では，声色との距離に基づいて推定する（式 (6)）ため，この付近に配置されていれば，正解と類似したスペクトル包絡が得られると考えられる．たとえば，SOLID を目標として，SOLID と VIVID と LIGHT の中間位置が推定された場合でも，SOLID に類似した音が合成できる可能性がある．以上の結果から，ユーザの声色変化を真似るための有効性が示唆された．

このようにし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の一部を図 10 に示す．実際の合成結果の一部はホームページ http://staff.aist.go.jp/t.nakano/VocaListener2/index-j.html で確認できる．

以上のように本実験では声色がフレーズごとにのみ変化する，比較的単純な正解データに基づいて評価を行った．今後は，実際の人間の歌唱を用いた場合や，様々な合成結果を用いた場合の評価に関して，システムの拡張を含めて研究の余地がある．

6. おわりに

本論文では，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．これによって，「このように歌ってほしい」という声色変化の制御を直感的かつ手軽に合成できる新しい手段を提供し，歌声合成手段の多様化につながる．また，VocaListener2 により，声色変化を手軽に制御して歌声合成でき，さらには音高・音量・声色変化の多様な観点から，歌唱の表情付けが行えるようになった．これによって，本人以外の多様な声質で手軽に歌唱曲を制作でき，歌声合成における声質や声色変化の多様化を可能にした．

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 1 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あげられる．本研究では声色変化の活用について 1 つの具体例を示したが，本実験で得られる声色空間は楽曲固有のものであり，曲ごとに構成し直す必要がある．ただし，予備的に行った実験では，別の曲に対しても似た傾向の空間が得られた．今後は異なる曲における声色空間の詳細な検討や，曲に依存しない空間の構成，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，さら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」の解明と，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．たとえば，VocaListener2 によって，音高や音量を真似た歌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，JST CREST 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] Kenmochi, H.: VOCALOID and Hatsune Miku Phenomenon in Japan, *Proc. 1st Interdisciplinary Workshop on Singing Voice* (*InterSinging 2010*), pp.1–4

(2010).
[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).
[5] Nakano, T. and Goto, M.: VocaListener: A Singing-to-Singing Synthesis System Based on Iterative Parameter Estimation, *Proc. 6th Sound and Music Computing Conference* (*SMC 2009*), pp.343–348 (2009).
[6] Abercrombie, D.: *Elements of General Phonetics*, Edinburgh University Press (1967).
[7] Laver, J.: *The Phonetic Description of Voice Quality*, Cambridge University Press (1980).
[8] 日本音響学会（編）：新版 音響用語辞典，コロナ社 (2003).
[9] Kenmochi, H. and Ohshita, H.: VOCALOID – Commercial Singing Synthesizer based on Sample Concatenation, *Proc. 8th Annual Conference of the International Speech Communication Association* (*INTERSPEECH 2007*) (2007).
[10] 粕谷英樹，木戸　博：声質の伝える情報とその関連量，日本音響学会 2011 年秋期講演論文集 1-8-9，pp.249–252 (2011).
[11] Sundberg, J.: *The Science of the Singing Voice*, Northern Illinois University Press (1987).
[12] Sundberg, J.: 歌声の科学，東京電機大学出版局 (2007).
[13] Villavicencio, F. and Bonada, J.: Applying Voice Conversion to Concatenative Singing-Voice Synthesis, *Proc. 11th Annual Conference of the International Speech Communication Association* (*Interspeech 2010*), pp.2162–2165 (2010).
[14] Türk, O., Büyük, O., Haznedaroglu, A. and Arslan, L.M.: Application of Voice Conversion for Cross-Language Rap Singing Transformation, *Proc. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP 2009*), pp.3597–3600 (2009).
[15] 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，Vol.2010-MUS-86, No.7, pp.1–10 (2010).
[16] 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（SP），Vol.110, No.297, pp.71–76 (2010).
[17] 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).
[18] Kawahara, H., Nisimura, R., Irino, T., Morise, M., Takahashi, T. and Banno, H.: Temporally Variable Multi-Aspect Auditory Morphing Enabling Extrapolation without Objective and Perceptual Breakdown, *Proc. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP 2009*), pp.3905–3908 (2009).
[19] 森勢将雅：歌声を混ぜるインタフェース「e.morish」，入手先 ⟨http://www.crestmuse.jp/cmstraight/personal/e.morish/⟩.
[20] Abe, M., Nakamura, S., Shikano, K. and Kuwabara, H.: Voice Conversion Through Vector Quantization, *Proc. 1988 International Conference on Acoustics, Speech and Signal Processing* (*ICASSP-88*), Vol.1, pp.655–658 (1988).
[21] Narendranath, M., Murthy, H.A., Rajendran, S. and Yegnanarayana, B.: Transformation of Formants for Voice Conversion Using Artificial Neural Networks, *Speech Communication*, Vol.16, pp.207–216 (1995).
[22] Desai, S., Raghavendra, E.V., Yegnanarayana, B., Black, A.W. and Prahallad, K.: Voice Conversion Using Artificial Neural Networks, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP 2009*), Vol.1, pp.3893–3896 (2009).
[23] 益子貴史，田村正統，徳田恵一，小林隆夫：HMM に基づく音声合成システムにおける MAP-VFS を用いた声質変換，電子情報通信学会論文誌 D，Vol.J83-D2, No.12, pp.2509–2516 (2000).
[24] Yamagishi, J. and Kobayashi, T.: Average-Voice-Based Speech Synthesis Using HSMM-Based Speaker Adaptation and Adaptive Training, *IEICE Trans. Information and Systems*, Vol.E90-D, No.2, pp.533–543 (2007).
[25] Stylianou, Y., Cappé, O. and Moulines, E.: Continuous Probabilistic Transform for Voice Conversion, *IEEE Trans. Speech and Audio Processing*, Vol.6, No.2, pp.131–142 (1998).
[26] Kain, A. and Macon, M.W.: Spectral Voice Conversion for Text-to-Speech Synthesis, *Proc. 1998 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP 1998*), Vol.1, pp.285–288 (1998).
[27] 望月　亮，大久保雅史，小林哲則：韻律情報を用いたスペクトル変換方式の検討，電子情報通信学会論文誌，Vol.J88-D-II, No.11, pp.2269–2276 (2005).
[28] Toda, T., Black, A. and Tokuda, K.: Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory, *IEEE Trans. Audio, Speech and Language Processing*, Vol.15, No.8, pp.2222–2235 (2007).
[29] 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).
[30] Villavicencio, F. and Maestre, E.: GMM-PCA based Speaker-Timbre Conversion on Full-Quality Speech, *Proc. 7th ISCA Workshop on Speech Synthesis* (*SSW-7*), pp.56–61 (2010).
[31] Doi, H., Nakamura, K., Toda, T., Saruwatari, H. and Shikano, K.: Esophageal Speech Enhancement based on Statistical Voice Conversion with Gaussian Mixture Models, *IEICE Trans. Information and Systems*, Vol.E93-D, No.9, pp.2472–2482 (2010).
[32] Nakamura, K., Toda, T., Saruwatari, H. and Shikano, K.: Speaking-Aid Systems Using GMM-based Voice Conversion for Electrolaryngeal Speech, *Speech Communication*, Vol.54, No.1, pp.134–146 (2012).
[33] Lee, C.-H. and Wu, C.-H.: Map-based Adaptation for Speech Conversion Using Adaptation Data Selection and Non-Parallel Training, *Proc. 9th International Conference on Spoken Language Processing* (*INTERSPEECH 2006*), pp.2254–2257 (2006).
[34] Mouchtaris, A., der Spiegel, J.V. and Muelle, P.: Nonparallel Training for Voice Conversion based on a Parameter Adaptation Approach, *IEEE Trans. Audio, Speech and Language Processing*, Vol.14, No.3, pp.952–963 (2006).
[35] Toda, T., Ohtani, Y. and Shikano, K.: Eigenvoice Conversion Based on Gaussian Mixture Model, *Proc. 9th In-*

ternational Conference on Spoken Language Processing (*INTERSPEECH 2006*), pp.2446–2449 (2006).
[36] Ohtani, Y., Toda, T., Saruwatari, H. and Shikano, K.: Improvements of the One-to-Many Eigenvoice Conversion System, *IEICE Trans. Information and Systems*, Vol.E93-D, No.6, pp.1589–1598 (2010).
[37] Saito, D., Yamamoto, K., Minematsu, N. and Hirose, K.: One-to-Many Voice Conversion Based on Tensor Representation of Speaker Space, *Proc. 12th Annual Conference of the International Speech Communication Association* (*INTERSPEECH 2011*), pp.653–656 (2011).
[38] Ohtani, Y., Toda, T., Saruwatari, H. and Shikano, K.: Non-parallel Training for Many-to-Many Eigenvoice Conversion, *Proc. 2010 IEEE International Conference on Acoustics Speech and Signal Processing* (*ICASSP 2010*), pp.4822–4825 (2010).
[39] Schröder, M.: Emotional speech synthesis: A review, *Proc. 7th European Conference on Speech Communication and Technology* (*Eurospeech 2001*), pp.561–564 (2001).
[40] Türk, O. and Schröder, M.: A Comparison of Voice Conversion Methods for Transforming Voice Quality in Emotional Speech Synthesis, *Proc. 9th Annual Conference of the International Speech Communication Association* (*Interspeech 2008*), pp.2282–2285 (2008).
[41] 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. Information and Systems*, Vol.E92-D, No.3, pp.489–497 (2009).
[42] Inanoglua, Z. and Young, S.: Data-driven Emotion Conversion in Spoken English, *Speech Communication*, Vol.51, No. 3, pp.268–283 (2009).
[43] Türk, O. and Schröder, M.: Evaluation of Expressive Speech Synthesis with Voice Conversion and Copy Resynthesis Techniques, *IEEE Trans. Audio, Speech and Language Processing*, Vol.18, No.5, pp.965–973 (2010).
[44] Iida, A., Campbell, N., Higuchi, F. and Yasumura, M.: A Corpus-based Speech Synthesis System with Emotion, *Speech Communication*, Vol.40, No.1-2, pp.161–187 (2003).
[45] Tsuzuki, R., Zen, H., Tokuda, K., Kitamura, T., Bulut, M. and Narayanan, S.S.: Constructing emotional speech synthesizers with limited speech database, *Proc. 8th International Conference on Spoken Language Processing* (*ICSLP 2004*), pp.1185–1188 (2004).
[46] 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).
[47] 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).
[48] 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).
[49] Onishi, M., Takahashi, T., Morise, M., Irino, T. and Kawahara, H.: Vowel-based Voice Conversion and Its Objective Evaluation, *Proc. 2008 RISP International Workshop on Nonlinear Circuits and Signal Processing* (*NCSP'08*), pp.275–278 (2008).
[50] Yoshida, Y., Nisimura, R., Irino, T. and Kawahara, H.: Vowel-Based Voice Conversion and Its Application to Singing-Voice Manipulation, *Proc. AES 35th International Conference: Audio for Games*, No.6 (2009).
[51] 廣瀬良文，釜井孝浩：PARCOR 係数補間による母音モーフィングに関する検討，電子情報通信学会技術報 音声（SP），Vol.110, No.297, pp.77–82 (2010).
[52] 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).
[53] 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).
[54] 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).
[55] 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).
[56] 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).
[57] Turk, G. and O'Brien, J.F.: Modelling with implicit surfaces that interpolate, *ACM Trans. Graphics*, Vol.21, No.4, pp.855–873 (2002).
[58] 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).

中野 倫靖 （正会員）

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，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．

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）．現在，産業技術総合研究所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．統計数理研究所客員教授，筑波大学連携大学院准教授，IPA 未踏 IT 人材発掘・育成事業 PM 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等，31 件受賞．