推薦論文

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的検討

河原英紀[†1] 生駒太一[†2] 森勢将雅[†2]
高橋徹[†1] 豊田健一[†3] 片寄晴弘[†4]

歌唱における演奏デザインの転写技術の確立を目指して，高品質音声分析変換合成システムSTRAIGHT に基づくモーフィングの研究を進めている．STRAIGHT を用いることにより，基本周波数，時間周波数表現，非周期性指標の 3 種類の物理パラメータのみから高度に自然な音声を合成することができる．ここでは，それらのパラメータと時間軸の変換関数および周波数軸の変換関数を併せた 5 種類の拡張パラメータを独立にモーフィングできるように拡張することで，歌手の声質と歌い回しを転写することを試みた．男子学生と男性のプロ歌手の歌唱を用いた実験刺激と，女性のプロ歌手 2 人の歌唱を用いた実験刺激の 2 種類の刺激が作成され，主観評価実験が行われた．実験結果は，まず，意図したとおりの転写が知覚されることと作成された刺激のほとんどが高い品質であることを示した．また，歌手の声質と歌い回しが簡単な物理パラメータの組合せで操作でき，ポピュラー音楽ではスペクトル情報に依存する声質が歌手の同定に大きく貢献していることが確認された．これらの結果に基づき，歌唱デザインインタフェースへの応用を論じた．

Proposal on a Morphing-based Singing Design Manipulation Interface and Its Preliminary Study

Hideki Kawahara,[†1] Taichi Ikoma,[†2] Masanori Morise,[†2]
Toru Takahashi,[†1] Ken'ichi Toyoda[†3] and Haruhiro Katayose[†4]

Investigations on singing voice morphing has been conducted to establish a design reuse framework based on a high-quality speech analysis, modification and resynthesis system STRAIGHT. STRAIGHT enables high-quality speech reconstruction using only three parameters; fundamental frequency, time-frequency representation, and time-frequency aperiodicity map. In this paper, an extension of STRAIGHT-based morphing, which enables individual control of morphing rate using five extended parameters (in addition to three parameters mentioned above, temporal axis mapping and frequency axis mapping functions were introduced), was implemented to test design reuse of singers' voice identity and their singing style. Subjective evaluations of two sets of manipulated samples were conducted. First set was generated from a male student and a professional male singer and the second set was generated from two professional female singers. Test results illustrated that intended reuse was perceptually verified and generated samples were generally in high-quality. The results also suggested that relatively simple combination of physical parameters enables intended reuse and identification of singer of the manipulated singing is mainly dependent on spectral information, for POP songs. Issues on singing style manipulation interface based on these findings were discussed.

1. はじめに

モーフィング技術には，物理パラメータと心理的属性との対応関係についての明示的な知識に依存せずに，2 つの試料の中間状態を容易に制作できることや，現実に存在しない効果を生み出すことができるという利点がある．これらの特長により，モーフィングはコン

†1 和歌山大学システム工学部
Faculty of Systems Engineering, Wakayama University
†2 和歌山大学大学院システム工学研究科
Graduate School of Systems Engineering, Wakayama University
†3（元）関西学院大学大学院理工学研究科
Graduate School of Science and Technology, Kwansei Gakuin University
†4 関西学院大学理工学部
School of Science and Technology, Kwansei Gakuin University

本論文の内容は 2007 年 3 月のシンポジウム「インタラクション 2007」にて報告され，同プログラム委員長により情報処理学会論文誌への掲載が推薦された論文である．

テンツデザインの有用なツールとして映像メディアの領域において広く実用に供されている．しかし，聴覚メディアの領域におけるモーフィングは，様々な要因により，広く利用されるには至っていない．

この事情を，伝説的なカストラートである Farinelli の歌声を映画[6] のために合成した例[7),8)] を手がかりとして見ることとする．カストラートはテノールからソプラノにわたる広大な声域を有するため，男性歌手と女性歌手の 2 人でそれぞれの声域を分担してカバーしなければならない．カストラートの声質は基本的にはカウンターテナーに類似しているため，まず女性歌手の声をモーフィングして近づけることが必要となる．また，声を若々しくし輝きを与えるために，スペクトルの全体的な形状の調整が行われる．しかもこれらの変換は，母音の種類や音高や発声の強さに依存するため，それらすべての組合せについてのデータベースを構築しなければならない．これらすべてを準備したうえで，母音のセグメントごとにこのデータベースを用いて目標とすべき周波数特性を選択し，入力音声を phase vocoder を用いて変換することにより，カストラートの歌声が合成される．なお，映画に使用できる品質を達成するためには，俳優の画像の口の動きに合わせて様々なパラメータを手作業で調整することが必要な部分もあったという．最終的には，この方法により総計で 39 分間に及ぶ高品質な歌唱が合成された．この例や最近の解説[4),23)] に見るように，高い品質の歌唱合成を本格的に行うためには，大量の素材の収集や音声生成の物理モデルにまで立ち入った処理が必要となる．

モデルにまで立ち入らない場合であっても，2 つの音の間のモーフィングのためには，波形そのものではなく音を表現する適切な物理パラメータによる表現上で補間することが必要となる．パラメータの間には補間特性の違いがあるため，たとえばスペクトルの概形と微細構造あるいは音源情報を分離し，独立に操作することとなる[3),22),24)]．しかし，自動的な手段によるそれらの情報の精密な分離はこれまで困難であり，簡便さと高い品質を両立させることは不可能であった．

この状況は，STRAIGHT[16)] と，それを用いたモーフィング技術の導入により変わりつつある[17),26),29)]．STRAIGHT は，聴覚の情景分析[5)] に代表される聴覚機能の生態学的理解に立脚して開発されたシステムである．STRAIGHT は，電気的音声処理の原点である channel VOCODER[9)] に遡り，開発された当時は二義的と見なされていた『音声を聴覚における情報表現と機能的に等価なパラメータ群に分解する』という側面を，現代の技術で追求することから生み出された[13),14)]．

STRAIGHT は，音声を基本周波数，スペクトル包絡，非周期性指標に分解する．注意深く相互の干渉を取り除かれて抽出された非負の実数で表されたそれらのパラメータは，操作による品質劣化の少ない音声変換を可能にする．STRAIGHT を，パラメータの値を変換しない単なる分析合成システムとして用いることもできる．その場合には，処理により波形が保存されないにもかかわらず，再合成された音声は，元の音声に匹敵する自然な聴覚的印象を与える[19)]．STRAIGHT に基づくモーフィングは，この特徴を利用して，パラメータおよびそれらの時間周波数座標を区分的一次関数を用いて補間することにより実現されている[17)]．

3 種類のパラメータだけから高い自然性を有する刺激連続体を作成できるというこの STRAIGHT に基づくモーフィングの特長[19),26),29)] は，たとえば，名歌手（故人）が最近のヒットチューンを歌ったらどうなるのかを聴いてみたいという願い[12)] や，自分の声でプロ並に歌いたい/それを残したいという願いを実現するための基盤を提供する．本論文では，そのような新しい能動的な音楽の愉しみ方や加工の上位概念である『歌唱デザインの転写』[11),12)] を実現するための第 1 歩として，異なった歌手による歌唱音声の声質と歌い回しのそれぞれを独立に操作する方法[28)] を提案し，予備的な検討を行った結果について報告する．

2. モーフィングによるデザインの転写

モーフィングは，2 つの事例となる試料が提供されたときにそれらの中間となる性質を有する試料を合成する操作として定義され，直感的なデザインインタフェースを構成できる可能性がある．しかし，この特長を生かすためには，知覚的属性に即した操作パラメータを用意する必要がある．ここでは，まず STRAIGHT に用いられているパラメータの性質とモーフィングへの応用について簡単に紹介し，その後，歌声操作のための拡張について述べる．

2.1 STRAIGHT の情報表現[14),16)]

STRAIGHT は，音声を基本周波数，スペクトル包絡，非周期性指標に分解する．再合成音声は，必要に応じて変形されたこれらのパラメータから高い周波数における群遅延にランダムな変動を加えられ非周期成分を加えられた音源信号と，最小位相応答システムとして実現されたフィルタを用いて生成される．

2.1.1 スペクトル包絡

このパラメータの分析では，まず，音声の基本周波

数に適応して時間長が変化する相補的時間窓を用いて，分析位置による変動のないパワースペクトルを求める．次いで，spline 関数の性質を利用した周波数方向の平滑化により，調波位置でのスペクトルのレベルを保存しかつ周期性の影響を除いたスペクトル包絡を抽出する．これらの処理によって基本周波数の情報がほぼ完全に除去されたスペクトル包絡は，周波数軸の伸縮や基本周波数の変換に起因する品質劣化が少ない音声の合成を可能とする．

2.1.2 基本周波数

STRAIGHT のスペクトル分析では，基本周波数に適応して窓関数が設計される．そのため，基本周波数の抽出誤りは品質劣化を引き起こす．音声の収録時に用いられる高域通過フィルタや低域に主要な成分を有する空調騒音および商用電源からの誘導雑音の影響は，周期性が弱くまた乱れることのある母音の開始および終了時の抽出誤りにつながる．これらの問題を解決するために，基本波および低次調波の瞬時周波数の情報と帯域ごとの正規化された自己相関とを併用し，さらに後処理を行う高精度の方法を開発した[15]．

2.1.3 非周期性指標

有声音であっても声門閉止が完全に行われない場合には，声門での乱流等に起因する非周期成分が含まれる．この現象は，女性の音声に典型的に認められる．男性の場合でも，有声摩擦音や弱い気息性の発声の場合には，顕著な非周期成分が含まれる[27]．これらを表現するために，STRAIGHT では非周期性を表す指標を用いている．指標は，帯域フィルタを通過するエネルギー全体と，その中に含まれている非周期成分のエネルギーの比（dB）として定義されている．帯域幅は，聴覚末梢系における周波数分解能を近似する ERB_N（Effective Rectangular Bandwidth）[21] に比例するように設定されている．実装では，基本周波数の瞬時位相を用いた時間軸の変換で仮想的に基本周波数を一定としたパワースペクトル[1),2)] の上側包絡と下側包絡の差から，非周期性指標を求めている[18]．

2.1.4 実　　装

STRAIGHT の実装には，科学技術計算用の環境である Matlab を用いた．STRAIGHT を構成するサブシステムは，基本周波数抽出用関数，非周期性指標抽出用関数，スペクトル包絡抽出用関数，合成音声作成用関数として実装されている．以下のモーフィングの実装も，Matlab 上でこれらの関数を用いて行った．

2.2 STRAIGHT に基づくモーフィングの拡張

STRAIGHT を用いることにより歌声は独立性の高い非負の実数のパラメータセットによって表現される．これによりモーフィングは事例間の線形補間によって容易に実装することができる．なお，外挿による負の値の発生を避けるため，実際のモーフィングは，対数変換したパラメータに対して行われた．ここで用いられているスペクトル包絡と音源情報という情報表現は，聴覚における属性との類似性が高く直感的に理解しやすい．以下では，まず，同一の楽譜に対する歌唱音声のモーフィングの手順を簡単に説明する．

2.2.1 スカラ値によるモーフィング

モーフィングは，2 つの歌唱のスペクトログラム（ここでは，各時刻のスペクトル包絡から作成される時間周波数表現）のそれぞれにおける対応する点（特徴点）が重なるように時間–周波数座標を変形させることから始まる．次いで，変形された時間–周波数座標の各点において，指定されたスカラ値（モーフィング率）に応じて 2 つの演奏のパラメータの値を補間/補外し，モーフィング率に応じて変形された時間–周波数座標にそれらの値を設定する．こうして作成されたパラメータを STRAIGHT の音声合成部に与えることで，目的とするモーフィング音声が合成される．

2 つの歌唱の時間軸と周波数軸の対応づけは，次のように行われる．まず，文献 17) と同様に，フォルマント軌跡の変曲点や音素境界等を手がかりとして，利用者の手作業により，特徴点がそれぞれの演奏の時間–周波数座標の上に設定される．次に，それぞれの演奏に付与された特徴点が重なり合うように，双一次変換により実装された関数を用いて時間–周波数座標が変形される．

2.2.2 モーフィング率の拡張とオブジェクト化

前項で説明したモーフィングでは，スペクトル包絡，基本周波数，非周期性指標と時間軸の変換関数および周波数軸の変換関数からなる計 5 個の拡張パラメータが，同一のモーフィング率で変換されていた．これは，パラメータ空間の 2 点として表現される事例の間を直線で結ぶ経路を表す．ここでモーフィング率 r を，媒介変数 λ を用いてパラメータ空間の各次元に対応する 5 個の係数（$r_1(\lambda)$, $r_2(\lambda)$, $r_3(\lambda)$, $r_4(\lambda)$, $r_5(\lambda)$）から構成されるベクトルに拡張することにより，2 点を結ぶ任意の経路を表現することができる．なお，実用上は，拘束条件（$r_k(0) = 0$, $r_k(1) = 1$, ただし $k \in \{1, 2, 3, 4, 5\}$）と単調性（$\lambda_1 \leq \lambda_2$ ならば $r_k(\lambda_1) \leq r_k(\lambda_2)$）を要請しておくと都合がよい．

これらの要請に柔軟に対応するために，モーフィングオブジェクトという概念を導入し，モーフィングを，2 つのモーフィングオブジェクトから 1 つのモーフィングオブジェクトを生成する演算として定義した．演

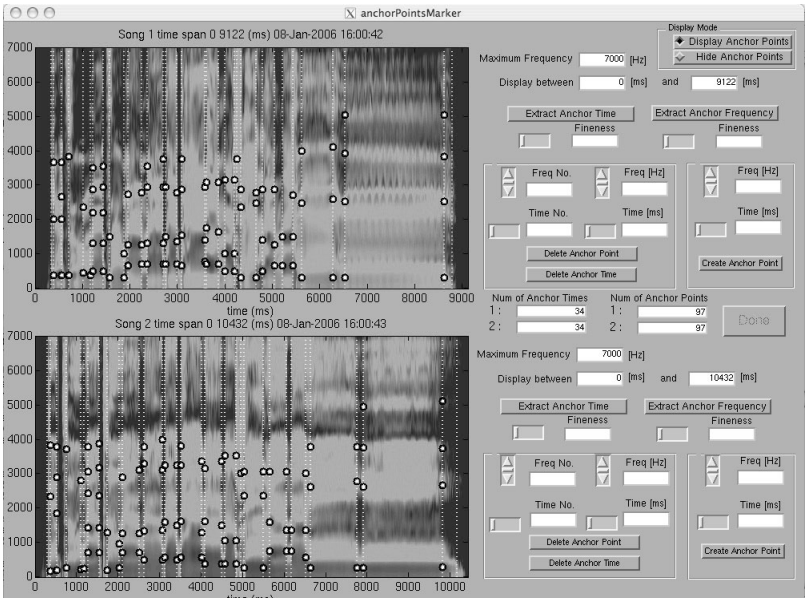

図 1 対応点付与作業支援ツール
Fig. 1 GUI tool for editing time-frequency markers.

算をこのように定義することで再帰的なモーフィングの適用が可能となり，2 つ以上の事例間のモーフィングへの拡張が容易となる．

実装では，モーフィングオブジェクトを Matlab の構造体として定義した．前述の 5 個のパラメータに加え，付与された特徴点の座標と分析条件やオブジェクトの履歴情報等を構造体のそれぞれのフィールドとした．

2.2.3 支援環境

モーフィングで用いる特徴点は，当初，研究者が実験目的に応じて注意深く設定することを想定していた[17]．しかし，音声生成および知覚についての基本的知識を欠く一般の利用者では適切な点の設定ができず，また，研究者にとっても，注意深い手作業には多くの時間を必要とするという問題がある．十分な信頼性を有する自動設定ツールがまだ存在していないため，ここでは，設定作業の支援に文献 28) の方法を利用した．

文献 28) の方法は，ガウス関数による多重解像度分析であるスケールスペースフィルタ[20] に基づいている．この方法では，まず特徴的な点の時刻の候補を，音量が大きく変化する点に相当するスケールスペースフィルタ出力の変曲点から求める．次いで，選択された時刻における周波数方向の特徴点の候補を，スペクトル包絡にスケールスペースフィルタを適用した出力の極大点から求める．求められる候補点は，用いるスケールスペースフィルタに含まれるガウス関数の標準偏差 σ に依存する．今回は，時間方向では 50 ms から 120 ms，周波数方向では 65 Hz から 215 Hz の範囲でいくつかの σ を用いた候補点の抽出を行い，**図 1** に示すような対話的ツールを用いて修正を行った．

図 1 の左側には，モーフィング対象の 2 人の歌手による歌声のスペクトログラム（STRAIGHT によるスペクトル包絡の系列の時間–周波数表現）が表示されている．白抜きの点は，設定された特徴点である．右側のパネルには，選択した特徴点や時刻を削除したり，周波数や時刻を編集したりするための GUI 要素が配置されている．作業の初期状態では，特徴点も時刻も過剰に抽出されている．ユーザは，スペクトログラムを手がかりとして，まず，子音や母音の開始・終了等に対応する時刻以外の不要な時刻を選択して削除する．次いで，それぞれの時刻における特徴点の中から，ホルマント等の重要な手がかりとなるものを残し，それ以外の不要な特徴点を削除する．作業は，2 つのスペクトログラムを比較して対応するものが残るように注意して進められる．また，必要に応じて周波数や時刻の数値を直接書き込んで変更することもできる．それぞれの歌唱音声に同じ個数の特徴となる時刻が設定され，それぞれの時刻において同じ個数の特徴点が設定されると，モーフィングの準備は完了する．この例では，「結末はまだ誰も知らない」という 9〜10 秒の一節に，最終的に 34 個の特徴となる時刻が設定され，合計 97 個の特徴点が設定された．

3. 声質と歌い回しのモーフィング

STRAIGHT を用いた実験により，聴覚は，音を生成している物体の形状やサイズの情報を無意識のうちに自動的に抽出しているらしいことが分かってきた[25]．歌唱音声の場合には，声道の形状は，それぞれの歌手の解剖学的な構造による拘束を受けているため，スペクトル包絡の張る空間での特定の歌手の歌唱音声の軌跡は，埋め込まれた小さな部分空間内にとどまると考えられる．STRAIGHT で求められるスペクトル包絡には，単一の声帯体積流波形のスペクトル形状も含まれるため，この部分空間は，声帯音源の個人性までを反映した，声質を総合的に表す特徴を表してもいる．したがって，ある歌手から別の歌手への声質の転写は，この部分空間どうしの写像を指定することにほかならない．これは，先に説明した 5 個の拡張パラメータから周波数軸の変換関数とスペクトル包絡だけを選択してモーフィングすることにより，歌手の声質の転写が実現できることを意味する．

一方，基本周波数の軌跡や音声を発するタイミングおよび音量は，調音器官の動特性による拘束を受けるものの，訓練により意図的に制御することが可能である．また，拘束の表れる調音器官の動特性の時定数の変動[27] と比較すると，基本周波数に深く関わっているピッチ知覚の時定数[21] ははるかに大きい．そのため，楽譜により基本周波数の平均値が拘束されている歌唱音声☆の聴取時に，基本周波数およびその軌跡から推定可能な発声器官の動特性の時定数が歌手の個人性を表すものとして利用されているとは考えにくい．これらの予備的な考察から，歌い回しのスタイルを転写するには，音素境界等に設定されている特徴点の相対時刻と基本周波数の軌跡とを歌手間で選択的にモーフィングすればよいことが示唆される．

なお，声質と歌い回しに関連するパラメータは，それぞれのグループ内で同一のものを用いることとするため，それぞれモーフィング率 r_{timbre}，r_{way} と表記することとする．

4. 声質と歌い回しの転写実験

前章で提案した仮説を検証するため，実際の歌唱音声を用いて主観評価実験を行った．以下に，用いた素材および手続きについて紹介する．

☆ 正確には，拘束されるのは知覚される属性であるピッチの系列である．

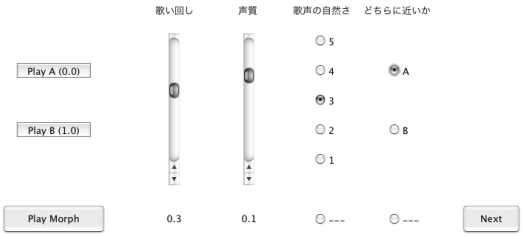

図 2 評価用アプリケーション
Fig. 2 GUI for subjective evaluation of morphed singing.

4.1 予 備 実 験

実験に用いた楽曲は，RWC 研究用音楽データベース[10] のポピュラー音楽に収録された楽曲 “So Long”（RWC-MDB-P-2001 No.64）のうち，サビ部分にあたる約 10 秒のフレーズである．音声試料は，上記の楽曲に収録されたスタジオミュージシャンによる歌声（歌声 A とする）と，同一曲を歌ったプロではない 20 代男性の歌声（歌声 B とする）である．歌声 B の歌声の収録において，歌唱者は楽曲の伴奏をヘッドホン（audio-technica ATH-A 1000）で聴きながら歌った．この歌声をマイク（SHURE SM57-LCE）により収録し，44.1 kHz，16 bit で記録した．なお，歌声 A はくせの強い歌い回しであったため，モーフィングの相手となる歌声 B の収録においては，なるべくピッチのゆらぎを少なくしてフラットに歌うよう指示した．

実験における刺激は，2 つの歌声を元にモーフィング率 r_{timbre}，r_{way} をそれぞれ −0.2 から 1.2 まで 0.2 ステップで変化させて生成したモーフィング音声 64 個と，歌声 A と歌声 B の計 66 個の歌声である．なお，歌声 A と歌声 B の特徴点は 7,000 Hz 以下に付与することとし，各時刻における特徴点の個数の上限を 5 とした．

4.1.1 評価用アプリケーション

被験者にはランダマイズされた刺激をヘッドホン（audio-technica ATH-A 1000）により提示し，各音声について以下のような 4 つの項目に対する評価を求めた．実験用に作成した評価用アプリケーションの外観を**図 2** に示す．被験者は左側の 3 個のボタンを用いることで，歌声 A，歌声 B，および刺激であるモーフィングされた歌声を何度でも聴くことができる．右下のボタンをクリックすることで評価が記録され，次の条件の刺激が提示される．

- 歌い回しのモーフィング率の評価
  刺激を聴取して感じた歌い回しのモーフィング率を左側のスライダを用いて評価する．スライダの下にフィードバック情報として表示される数値は −0.4 から 1.4 の間で 0.1 刻みとした．

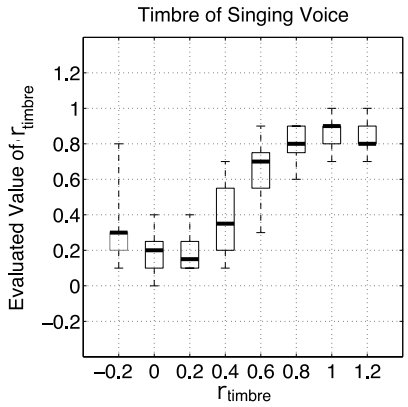

図 3 声質のモーフィング率の評価
Fig. 3 Timber rating results for r_{timbre} manipulation.

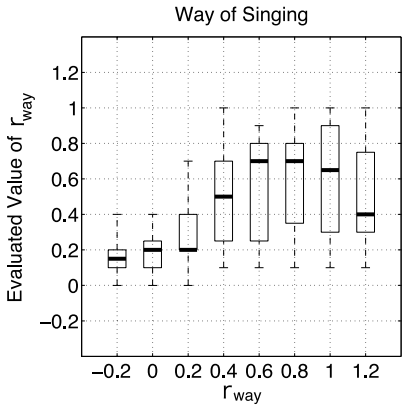

図 4 歌い回しのモーフィング率の評価
Fig. 4 Style rating results for r_{way} manipulation.

- 声質のモーフィング率の評価
  刺激を聴取して感じた声質のモーフィング率を右側のスライダを用いて評価する．スライダの下にフィードバック情報として表示される数値は −0.4 から 1.4 の間で 0.1 刻みとした．
- 歌声の自然性の評価
  刺激の自然性を「非常に自然」から「非常に不自然」までの 5 段階で評価する．
- 歌唱者の判別
  刺激が歌声 A と歌声 B のいずれにより近いと感じられるかを 2 択で選択する．

歌唱音声のモーフィングは，たとえば「もう少し○○さん風の歌い回しにしたい」「もう少し○○さんの声に似せたい」「○○さんの感じを誇張して加えたい」という要求を実現するための手段となる．このような要求を，任意の値を連続的に設定できるスライダの操作にマッピングする際の知見を得ることを狙い，評価用アプリケーションの GUI を決定した．今回の設定では，負のモーフィング率と 1 を超えるモーフィング率が誇張する操作に対応している．

4.1.2 モーフィング率の評価結果

3 人の被験者によるモーフィング音声の声質のモーフィング率の評価結果を箱ひげ図を用いて**図 3** に示す．横軸は，操作に用いた声質のモーフィング率 r_{timbre} を表し，縦軸は，被験者によって評価されたモーフィング率を示す．箱の外に飛び出した横棒（ひげ）は，評価の最大値最小値を示し，箱の下辺と上辺とは，それぞれ累積分布の 25%点と 75%点を示す．太線で示した箱の中の横棒は，評価値の平均値を表す．モーフィングが内挿となっている領域では，操作したモーフィング率と主観的に評価されたモーフィング率は，ほぼ単調に対応している．しかし，モーフィングが外挿となっている部分では，この単調性は崩れている．また，操作した範囲が −0.2 から 1.2 であったにもかかわらず，すべての評価値が 0 から 1 の範囲に入っている．

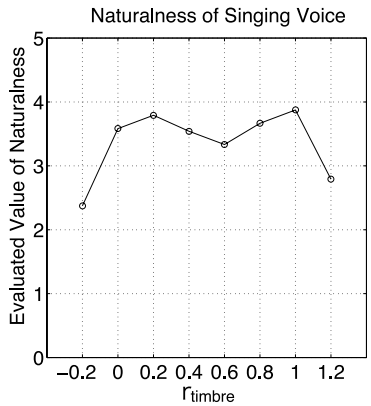

図 5 声質のモーフィング率と歌声の自然性
Fig. 5 Naturalness rating for r_{timbre} manipulation.

図 4 に，歌い回しの評価結果を示す．横軸は，操作に用いた歌い回しのモーフィング率 r_{way} を表し，縦軸は，被験者によって評価されたモーフィング率を示す．箱ひげ図の表現方法は図 3 と同じである．歌い回しの評価は，声質と比較すると評価値が広い範囲に分布している．評価値の個人差も大きく，声質と比較すると歌い回しは評価が困難であるものと考えられる．

これらの結果は，モーフィング操作についての十分な説明が与えられていない場合には，被験者が 0 から 1 の範囲を超える値の意味のイメージを持てないことを示唆する．同様の傾向は文献 29) にも認められる．

4.1.3 自然性の評価結果

図 5 に声質のモーフィング率に対する自然性の評価値の平均値，**図 6** に歌い回しのモーフィング率に対する自然性の評価値の平均値を示す．

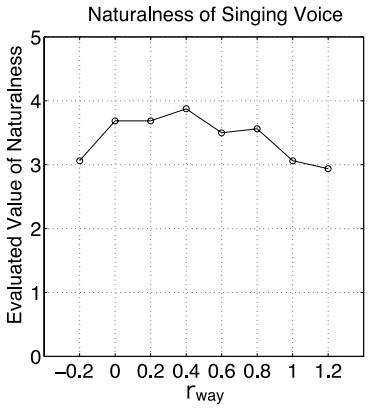

図 6 歌い回しのモーフィング率と歌声の自然性
Fig. 6 Naturalness rating for r_{way} manipulation.

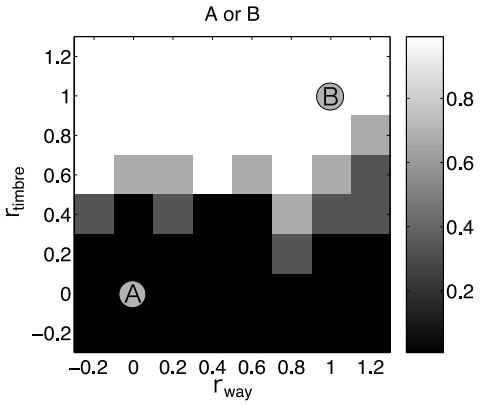

図 7 歌声の判別（A と判断された場合を 0，B と判断された場合を 1 とし，被験者の評価値の平均値を表している）
Fig. 7 Singer identification map for r_{timber} and r_{way} manipulations. Dark color represents singer A and white color represents singer B.

モーフィングの元試料となる 2 つの歌声の自然性の評価値の平均はいずれも 4.5 であった．単なる分析合成による歌声の自然性の評価値は 3.8 であり若干の低下が認められたが，内挿を行った場合のモーフィング音声の自然性は，この単なる分析合成と同程度であった．しかし，外挿を行った場合のモーフィング音声では自然性が大きく低下した．この傾向は，声質のモーフィングの場合に顕著である．実験後の被験者のコメントには，外挿を行ったモーフィング歌声について「声質が荒れているように感じた」とするものがあった．

4.1.4 歌声の判定

被験者によるモーフィング音声の歌唱者の判別結果を**図 7** に示す．横軸は歌い回しのモーフィング率 r_{way}，縦軸は声質のモーフィング率 r_{timbre} に対応している．なお図中の◯印の A，B は，それぞれ歌手 A の分析合成音声と歌手 B の分析合成音声と同じものとなる．評価値は，A ならば 0，B ならば 1 とし，グラフの色は被験者 3 人による評価値の平均値を表している．グラフの黒い部分ほど歌声 A，白い部分ほど歌声 B と判断されたことを表している．

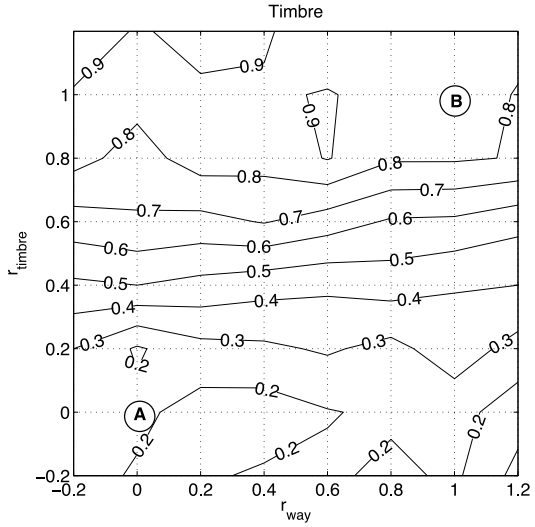

図 8 声質のモーフィング率の評価値のマップ
Fig. 8 Timber rating map for r_{way} and r_{timbre} manipulation.

この結果から，歌唱者の判別においては，歌い回しよりも声質のほうが支配的であることが分かる．ただし，実験のサンプルに用いた 2 つの歌声の歌い回しが類似していたという可能性も否めない．そこで，歌唱者と被験者を代え次の実験を企画した．

4.2 実　　験

ここでは，2 人のプロの女性歌手による 8 秒間の歌唱音声を用いて学生を中心とする 8 人の被験者による実験を行った．実験は，各被験者について 2 回実施されており，それぞれの刺激は計 16 回の評価を受けた．曲は，著作権処理の問題がクリアされた典型的なポピュラー音楽風の新曲（Love affair）を用いた．被験者による実験の手続きは予備実験と同様である．ただし，被験者は他の研究において音声モーフィングを用いており，モーフィング率の意味についての理解と外挿のイメージをすでに有している．

4.2.1 モーフィング率の評価結果

今回の実験では実験回数が多く，分布を詳細に見ることができる．まず，操作に用いた声質と歌い回しのモーフィング率と被験者によるそれぞれの率の評価結果との対応関係を調べた．**図 8** に声質の評価の平均値を，**図 9** に歌い回しの評価の平均値を横軸を歌い回しのモーフィング率，縦軸を声質のモーフィング率とする平面上の等高線を用いて示す．声質の評価結果

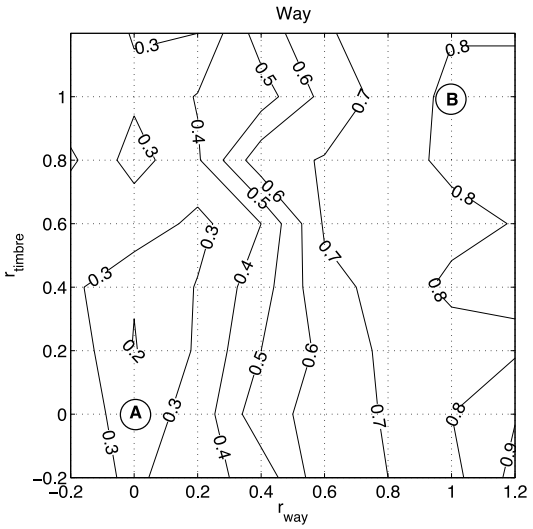

図 9 歌い回しのモーフィング率の評価値のマップ
Fig. 9 Style rating map for r_{way} and r_{timbre} manipulation.

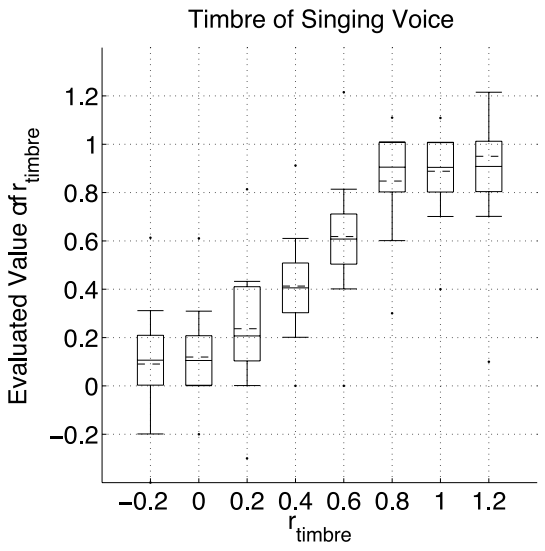

図 10 声質のモーフィング率の評価
Fig. 10 Timber rating results for r_{timbre} manipulation.

の等高線は歌い回しの軸に平行であり，歌い回しの評価結果の等高線は声質の軸に並行である．これらの結果は，モーフィングに用いたパラメータの組合せが知覚される声質と歌い回しによく対応していることを意味する．

図 10 と**図 11** にそれぞれ声質と歌い回しのモーフィング率の評価と対応する操作量との関係を示す．横軸と縦軸は，予備実験の図と同じである．ここでは，黒点で表した評価の最大値と最小値に加え，ヒゲを用いて 10%および 90%の点を表し，箱内の実線で分布の

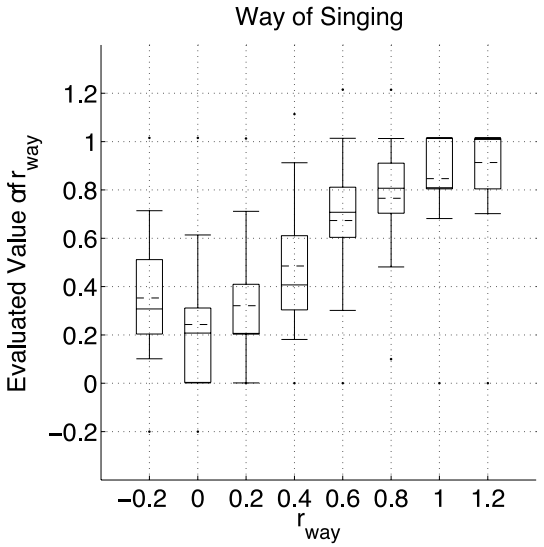

図 11 歌い回しのモーフィング率の評価
Fig. 11 Style rating results for r_{way} manipulation.

中央値を，破線で平均値を表すこととした．

図 10 に示した声質の評価では，外挿に相当する評価を示す被験者もおり，評価値は，より広い範囲に分布している．しかし，予備実験と同様に，物理的な操作量が外挿にあたる領域では，評価値が飽和する傾向が認められた．

図 11 に示す歌い回しの評価では，予備実験と比較すると，より分布が狭いことが分かる．今回の 2 人のプロ歌手は，声質も歌い回しも，予備実験の男性と比較すると違いがはっきりとしているため，判定がより容易であった結果であると思われる．ただし，実験終了後のインタビューでは，被験者の多くは，歌い回しは分かりにくく回答に迷ったとコメントしていた．得られた評価値も，声質の判定と比較すると外挿領域での飽和傾向が強く，また，相対的に分布が広がっていることが分かる．

4.2.2 自然性の評価結果

図 12 に，自然性の評価結果の平均値を，横軸を歌い回しのモーフィング率，縦軸を声質のモーフィング率とする平面上の等高線を用いて示す．元の歌唱音声の自然性の評価値の平均は，歌手 A が 4.5，歌手 B が 4.56 であった．モーフィングされた歌唱音声の自然性の評価値の平均は，1.25 から 4.25 の範囲に分布しており，元の歌唱音声に匹敵するものがあることが分かる．特に，2 つの事例を結ぶ対角線上の条件で自然性が高い．

図 12 の縦方向の変化と横方向の変化を比較することにより，外挿による品質劣化が声質のモーフィング

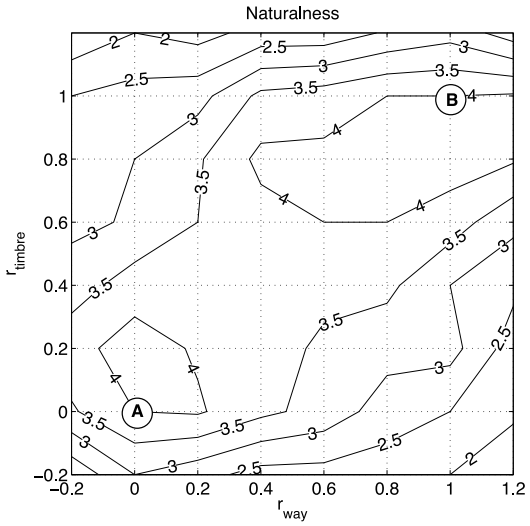

図 12 歌い回しと声質のモーフィングによる歌声の自然性
Fig. 12 Naturalness rating for r_{way} and r_{timbre} manipulation.

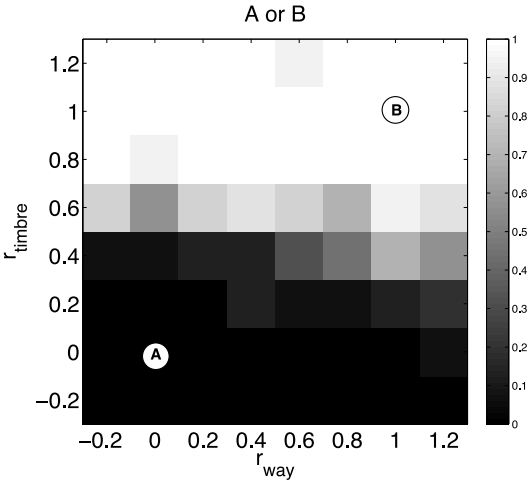

図 13 歌声の判別（A と判断された場合を 0，B と判断された場合を 1 とし，被験者の評価値の平均値を表している）
Fig. 13 Singer identification map for r_{timber} and r_{way} manipulations. Dark color represents singer A and white color represents singer B.

において顕著であることが分かる．この結果は，対数スペクトルの線形補外という単純な方法による外挿がスペクトルの自然な構造を壊すものであることを意味する．

実際，外挿にあたる部分では，2 つの事例の対数スペクトルの差が強調されるため，スペクトル上に余分なピークが出現する等の品質劣化につながる問題が生じている．

4.2.3 歌声の判定

図 13 にモーフィングされた歌声の歌唱者の判定結果を示す．予備実験と同様に，被験者は主に声質によって歌唱者を判定していることが分かる．予備実験とは異なり，今回の 2 人のプロ歌手による演奏は，歌い回しの判定が比較的容易であった．それにもかかわらず，予備実験と同様に，被験者は声質に基づいて歌唱者を判定した．今後，より多くの実験によって検証することが必要ではあるが，歌唱者の判定では声質が支配的であると結論づけてよいであろう．

これらの結果は，収録された歌唱のポストプロダクションにおいて歌手の声質の転写や歌い回しの転写を可能にするインタフェースを設計する際の有用な知見を与える．また，歌手の判定がほぼ声質のみに依存していることは，インタラクティブなリアルタイムの歌手変換が実現可能であることを意味する．

5. おわりに

歌唱音声の歌手の声質や歌い方の自由な操作を目的に，STRAIGHT に基づくモーフィングを拡張し，パラメータ操作と知覚印象との関係を調べた．実験結果は，基本周波数と時間軸の操作により歌い回しが，スペクトル包絡および非周期性指標と周波数軸の操作により声質が選択的に変換されることと，また，ポピュラー音楽では歌手の個人性の判断が主に声質により行われることを示唆するものであった．これらの結果は，モーフィングに基づく歌唱デザインインタフェースが，歌手の声質と歌い回しを直感的に操作するための有用な手段となりうることを示すものである．

このインタフェースを実用的なものとするためには，多くの課題の解決が必要である．まず，手作業の必要な特徴点の付与が実用する際の大きな障害となる．特徴点設定の自動化や特徴点の設定そのものを不要とするモーフィング手法の開発が急務である．また，外挿領域での自然性の大きな劣化は，物理パラメータの時間周波数座標における値そのものを補間するという現在のモーフィングが，声質と歌い回しの本質をとらえていないことを意味する．分析合成を行うことだけで生ずる品質の劣化も，コンテンツのポストプロダクションへの本格的な応用では大きな問題となる．モーフィングを行う際に目標となる事例が必要であることは，長所でもあり短所でもある．「加工対象と同一の内容（旋律，歌詞）の目標が必要」という条件を緩和することが課題となる．

これらの課題を解決してゆくことを通じて，本論文の最初で紹介した表現とデザインに関する願いを実現する様々な技術とツールを創り出してゆきたい．

謝辞 本研究は，科学技術振興機構 CrestMuse プ

ロジェクトの支援を受けた．

参 考 文 献

1) Abe, T., Kobayashi, T. and Imai, S.: Harmonics estimation based on instantaneous frequency and its application to pitch determination, *IEICE Trans. Information and Systems*, Vol.E78-D, No.9, pp.1188–1194 (1995).
2) Abe, T., Kobayashi, T. and Imai, S.: The IF Spectrogram: A New Spectral Representation, *Proc. ASVA-97*, Tokyo, pp.423–430 (1997).
3) 坂野秀樹，武田一哉，鹿野清宏，板倉文忠：包絡と音源の独立操作による音声モーフィング，電子情報通信学会誌，Vol.J81-A, No.2, pp.261–268 (1998).
4) Bonada, J. and Serra, X.: Synthesis of the singing voice by perfromance sampling and spectral models, *IEEE Signal Processing Magazine*, Vol.24, No.2, pp.67–79 (2007).
5) Bregman, A.S.: *Auditory Scene Analysis*, MIT Press, Cambridge, MA (1990).
6) Corbiau, G.: Farinelli il castrato (1994). （映画：邦題：カストラート）.
7) Depalle, P., Garcia, G. and Rodet, X.: A Virtual Castrato (!?), *Proc. 1994 International Computer Music Conference*, Aarhus, Denmark, pp.357–360 (1994).
8) Depalle, P., Garcia, G. and Rodet, X.: The recreation of a castrato voice, Farinelli's voice, *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.242–245 (1995).
9) Dudley, H.: Remaking speech, *J. Acoust. Soc. Am.*, Vol.11, No.2, pp.169–177 (1939).
10) 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).
11) 片寄晴弘，後藤真孝：音楽のデザイン転写技術の開発にむけて — CrestMuse プロジェクトの「価値」創出視点からの紹介，人工知能学会近未来チャレンジ，pp.1D1–4 (2006).
12) 河原英紀，片寄晴弘：高品質音声分析変換合成システム STRAIGHT を用いたスキャット生成研究の提案，情報処理学会論文誌，Vol.43, No.2, pp.208–218 (2002).
13) 河原英紀：Vocoder のもう一つの可能性を探る—音声分析変換合成システム STRAIGHT の背景と展開，日本音響学会誌，Vol.63, No.8, pp.442–449 (2007).
14) Kawahara, H.: STRAIGHT, Exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds, *Acoust. Sci. & Tech.*, Vol.27, No.6, pp.349–353 (2006).
15) Kawahara, H., de Cheveigné, A., Banno, H., Takahashi, T. and Irino, T.: Nearly Defect-free F0 Trajectory Extraction for Expressive Speech Modifications based on STRAIGHT, *Interspeech'05*, Lisboa, pp.537–540 (2005).
16) Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A.: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction, *Speech Communication*, Vol.27, No.3-4, pp.187–207 (1999).
17) Kawahara, H. and Matsui, H.: Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation, *Proc. ICASSP 2003*, Vol.I, Hong Kong, pp.256–259 (2003).
18) Kawahara, H., Katayose, H., de Cheveigné, A. and Patterson, R.D.: Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity, *Proc. Eurospeech'99*, Vol.6, pp.2781–2784 (1999).
19) Matsui, H. and Kawahara, H.: Investigation of Emotionally Morphed Speech Perception and its Structure using a High Quality Speech Manipulation System, *Proc. Eurospeech'03*, Geneva, pp.2113–2116 (2003).
20) Mayer, H. and Steger, C.: A New Approach For Line Extraction and its Integration in a Multi-Scale, Multi-Abstraction-Level Road Extraction System, *Mapping Buildings, Roads and other Man-Made Structures from Images*, Leberl, F., Kalliany, R. and Gruber, M. (Eds.), pp.331–348, Wien, R. Oldenbourg Verlag (1996).
21) Moore, B.C.J.: *An introduction to the psychology of hearing: 5th edition*, Academic Press (2003).
22) Pfitzinger, H.R.: Unsupervised speech morphing between utterances of any speakers, *Proc. 10th Australian International Conference on Speech Sciences & Technology*, Sydney Australia, pp.545–550 (2004).
23) Rodet, X.: Synthesis and processing of the singing voice, *Proc.IEEE MPCA-2002*, Leuven, Balgium, pp.99–108 (2002).
24) Slaney, M., Covelle, M. and Lassiter, B.: Automatic audio morphing, *ICASSP 1996*, Vol.2, Atlanta, pp.1001–1004 (1996).
25) Smith, D.R., Patterson, R.D., Turner, R., Kawahara, H. and Irino, T.: The processing and perception of size information in speech sounds, *J. Acoust. Soc. Am.*, Vol.117, No.1,

pp.305–318 (2005).
26) 曽我部優子，筧 一彦，河原英紀：感性情報に曖昧さがある場合の音声の心理的評価とその物理特性，聴覚研究会資料，H-2003-14, pp.77–82 (2005).
27) Titze, I.R.: *Principles of voice production*, Prentice Hall (1994).
28) 豊田健一，片寄晴弘，河原英紀：STRAIGHT による歌声モーフィングの初期的検討，情報処理学会研究報告，Vol.2006-MUS-64 (2006).
29) Yonezawa, T., Suzuki, N., Mase, K. and Kogure, K.: HandySinger: Expressive Singing Voice Morphing using Personified Hand-puppet Interface, *Proc. NIME2005*, Hamamatsu, pp.121–126 (2005).

(平成 19 年 4 月 3 日受付)
(平成 19 年 7 月 3 日採録)

推 薦 文

歌唱音声をモーフィングするインタフェースの開発というユニークな研究報告である．主観評価実験により，歌手の個人性知覚には歌い回しよりも声質の方が強く寄与することを明らかにし，科学的知見として興味深い．シンポジウム予稿集用のカメラレディ原稿をもとに 47 名のシンポジウムプラグラム委員による審査・投票を行い，大変評価が高かったため，論文誌に推薦した．
（インタラクション 2007 プログラム委員長 角 康之）

河原 英紀（正会員）
1972 年北海道大学工学部電子工学科卒業．1977 年同大学大学院工学研究科博士課程修了，工学博士．同年電電公社武蔵野電気通信研究所入所．1992 年 ATR 人間情報通信研究所第一研究室長．1997 年より和歌山大学システム工学部教授．聴覚機能の数理的解明と工学的表現の研究に従事．1997 年音響学会佐藤論文賞．2000 年 EURASIP 最優秀論文賞受賞．IEEE，ASA，日本音響学会，神経回路学会，認知科学会各会員．

生駒 太一
2006 年和歌山大学システム工学部卒業．現在，同大学大学院システム工学研究科博士前期課程 2 年．歌唱モーフィングの研究に従事．

森勢 将雅
2004 年和歌山大学システム工学部卒業．2006 年同大学大学院システム工学研究科博士前期課程修了．現在，同大学院博士後期課程に在学．音声音響信号処理に関する研究に従事．日本学術振興会特別研究員 DC1．2007 年電気通信普及財団テレコムシステム学生技術賞受賞．日本音響学会，電子情報通信学会各会員．

高橋 徹（正会員）
1996 年名古屋工業大学知能情報システム学科卒業．2004 年同大学大学院電気情報工学博士課程修了．同年和歌山大学システム工学部産学官連携研究員．博士（工学）．音声分析・合成・歌声合成の研究，特に音声モーフィング・音声テクスチャマッピングに基づく声質変換に関する研究に従事．音声分析変換合成システム STRAIGHT の開発に参加．電子情報通信学会，日本音響学会各会員．

豊田 健一
2006 年関西学院大学大学院理工学研究科修士課程修了．工学修士．現在，NTT コムウェア（株）勤務．

片寄　晴弘（正会員）

1991 年大阪大学大学院基礎工学研究科博士課程修了．工学博士．イメージ情報科学研究所，和歌山大学を経て，現在，関西学院大学理工学部教授．ヒューマンメディア研究センターセンター長．音楽情報処理，感性情報処理，HCI の研究に従事．科学技術振興機構さきがけ研究 21「協調と制御」領域研究者．科学技術振興機構 CREST「デジタルメディア（略称）」領域 CrestMuse プロジェクト代表研究者．電子情報通信学会，人工知能学会各会員．