Vol.2012-MUS-96 No.6
2012/8/9

声質空間上での変換に基づく歌声らしさの転写に関する検討

齋藤 大輔[1,a)]　石原 達馬[1]　橘 秀幸[1]　亀岡 弘和[1,2]　嵯峨山 茂樹[1]

概要：本稿では，特定話者の声に存在する歌声らしさを抽出し，任意の話者モデルに付加することで，歌声モデルを構築する手法を提案する．近年の話声の音声合成技術の発展に伴って，歌声合成技術は飛躍的に発展しているが，コーパスベースの手法における歌声合成モデルの個人性は，そのモデルを構築するためのデータベースに大きく依拠する．一方，話者適応技術を応用することで，少量の音声データを用いて対象歌手の歌声モデルを構成することも考えられるが，その音声データのうち不要な情報も含めて適応してしまう可能性がある．本研究では，音声データに含まれる声道特性に起因する情報のうち，「歌声らしさ」のような部分的情報に着目して，それを適切に転写する手法を検討した．本研究では任意話者声質変換で用いられる重みベクトル/重み行列の特徴量空間を声質空間と捉える．この空間上で同一話者の話声及び歌声が記述され，その変換関係によってこの話者の歌声らしさが表現される．この変換を別話者の話声モデルに適用することで，歌声らしさの転写を実現する．歌声の声質変換実験を通して，提案法による歌声らしさの転写が可能であることを示した．提案法を用いることで話声と歌声のモデル・データを相互に柔軟に利用することが可能となる．

キーワード：歌声声質変換, 任意話者声質変換, 声質空間, 歌声らしさ, 差分転写

Transfer of Singing Voice Likeliness Using Conversion on Voice Quality Space

Daisuke Saito[1,a)]　Tatsuma Ishihara[1]　Hideyuki Tachibana[1]　Hirokazu Kameoka[1,2]
Shigeki Sagayama[1]

Abstract: This paper describes a novel approach to construct a singing voice model for an arbitrary singer by transferring and appending singing voice likeliness from another speaker's voice. In corpus-based methods in speech or singing voice synthesis, speaker/singer individuality of a constructed voice model depends on the database used for construction of the model. Although speaker adaptation techniques enable the construction of a target singer's model by a small amount of speech data, over-adaptation including unnecessary information could occur. In this study, we propose a method to select some focused information included in speech data of a speaker, such as "singing voice likeliness," and to transfer the extracted information to another speaker's model properly. In the proposed method, a weight feature space used in arbitrary speaker conversion is regarded as a voice quality space. Speaking and singing voices of the speaker are respectively described in this space, and the singing voice likeliness of the speaker is represented as conversion between these two voices in the space. This conversion is applied to another speaker's model, and then transfer of singing voice likeliness is achieved. Voice conversion experiments using singing voices show that the proposed method properly transfers singing voice likeliness. The proposed method enables us to utilize both singing and speaking voices/models flexibly.

Keywords: Singing Voice conversion, Arbitrary Speaker Conversion, Voice Quality Space, Singing Voice Likeliness, Difference Transfer

1. はじめに

歌唱音声を人工的に生成する歌声合成システムは，エンターテインメント応用をはじめとして多様な可能性を持っている．話声における音声合成技術が発展していく中で，歌声合成技術も大きく進歩してきている．近年の音声合成技術の主流である素片接続型音声合成および HMM 音声合成は，大規模な音声コーパスを基盤とした技術である．使用する音声コーパスの質の向上及び量の増大に伴って，音声合成の品質は飛躍的に向上している．しかし，コーパスベースの手法における音源（接続素片）やモデルの構築は，そのために用いるデータベースに強く依存している．構築される声のモデルは使用するデータベースの個人性やスタイルを反映したものになる．例えば，話声の音声データしか存在しない状況において，歌声音声合成のための音源やモデルを構築するのは容易ではない．一方で，近年では話声・歌声を問わず多種多様な音声データを利用可能となっている．これらの豊富なデータから部分的に必要な情報を組み合わせて利用することができれば，ユーザによる柔軟なモデル制御や自由な楽曲製作支援につながる．例えば個人性はある話者 A の話声データから推定し，歌い方の癖などを歌手 B のものを表現するといったことが可能になれば，より自由な創作も実現できる．

ユーザの音声を入力して歌声合成を実現する手法として，VocaListener [1], [2] や SingBySpeaking [3] がある．VocaListener ではユーザの歌唱音声を入力として，その歌唱の音高と音量を真似るように既存の音声合成ソフトウェアの合成パラメータを調整できる．また VocaListener2 では上記に加えて，ユーザの声色変化も転写の対象としている．一方 SingBySpeaking では，ユーザが歌詞を朗読した音声を入力として，これを歌声特有の音響特徴を制御することで，話声から歌声への変換を実現する．これらの手法はユーザの入力音声から着目する情報を抽出し利用していると考えることができる．しかしこれらの手法では，ユーザの入力が歌唱内容を制約することになる（歌詞やメロディなど）．

一方，統計的な音響モデルを用いた歌声合成システムでは，歌詞を入力とした lyric-to-singing synthesis を実現でき，また話者適応技術を利用することで少量の音声データからその特徴を反映したモデルを構築することも可能である [4]．しかし適応に用いるデータのうち，個人性やそのスタイルもすべて含めて適応されるため，特定の情報のみを適応したい（例えば歌声らしさのみの適応など）といった目的に対しては，必ずしも適合しない．

本研究では，主に音声データに含まれる声道特性に起因する情報に着眼して，それらのうち「歌声らしさ」のような部分的情報の特徴を抽出・転写する手法を検討する．これらの情報を抽出するには，着目する情報以外，例えば音声データの話者性に関する情報の影響を適切に取り除く必要がある．本研究では任意話者声質変換で用いられる重みベクトルや重み行列に着目し，この特徴量空間を声質を定量的に表す声質空間であると考える [5], [6]．任意話者声質変換では，話声の話者性を変換することを目的とし，重みの特徴量空間の一点が特定の話者を表現すると考える．一方，本研究では，同一の話者であっても話声と歌声といった発声の違いによって，声質空間における記述が変化すると仮定し，この変化に着目する．同一話者内での話声と歌声の違いを声質空間上での変換として捉え，この変換を異なる話者の話声特徴に対して適用することで，「話声と歌声の違い」の転写を実現する．

以下，2 章において，任意話者声質変換法を利用した声質空間の構築について述べる．3 章で，構築した声質空間上での変換に基づく声色転写法について述べる．4 章において提案法が適切に歌声らしさを転写できるかに関する実験的評価について述べ，最後に 5 章で，本稿の結論と今後の課題について述べる．

2. 任意話者声質変換を用いた声質空間の構築

2.1 固有声に基づく混合正規分布モデル

本章では，一対多固有声変換法 (Eigenvoice conversion; EVC) 及びテンソル表現に基づく任意話者声質変換 (Tensor-based arbitrary speaker conversion; TASC) について概説し，これらの話者表現を用いた声質空間構築について説明する [5], [6].

声質変換は，入出力の対応関係を記述する変換モデルに基づいて，任意の文に対して入力音声の声質を所望の声質へ変換する技術である．声質変換は，広義には異なる二つの特徴量空間のマッピング技術と考えられ，テキスト音声合成における話者性の制御をはじめとして [7]，雑音環境下音声の音声強調や身体運動から音声への変換など多岐にわたる応用が検討されている [8], [9]．しかし，変換モデルの構築の際には，基本的に同一発話内容の入出力音声対からなるパラレルデータを用いる必要がある．この問題を解決するために，多数の話者の音声データをより効果的に用いて，入出力話者のパラレルデータなしに任意話者を対象とした声質変換を実現するのが EVC である．

EVC では，参照話者と多数の事前収録話者との間の複数のパラレルデータを用いて，固有声に基づく混合正規分布モデル（EV-GMM）を構築する．今，参照話者の音響特

[1] 東京大学 大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113–0033, Japan
[2] NTT コミュニケーション科学基礎研究所
Communication Science Laboratories, NTT Corporation, 3-1 Wakamiya, Morinosato, Atsugi, Kanagawa 243-0198, Japan
[a] dsaito@hil.t.u-tokyo.ac.jp

徴量を $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \boldsymbol{\Delta x}_t]^\top$，$s$ 番目の事前収録話者の音響特徴量を $\boldsymbol{Y}_t^{(s)} = [\boldsymbol{y}_t^{(s)^\top}, \boldsymbol{\Delta y}_t^{(s)^\top}]^\top$ と表す．ただし $^\top$ は転置を表す．ここで，音響特徴量は D 次元の静的および動的特徴量を結合した $2D$ 次元の音響特徴量となる．参照話者と事前収録話者の結合確率密度は，EV-GMM として以下のようにモデル化される．

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}^{(s)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^{(s)^\top}]^\top; \boldsymbol{\mu}_m^{(Z)}(\boldsymbol{w}^{(s)}), \boldsymbol{\Sigma}_m^{(Z)}) \quad (1)$$

$$\boldsymbol{\mu}_m^{(Z)}(\boldsymbol{w}^{(s)}) = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{B}_m \boldsymbol{w}^{(s)} + \boldsymbol{b}_m^{(0)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \quad (2)$$

ここで $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は，平均ベクトルを $\boldsymbol{\mu}$，分散共分散行列を $\boldsymbol{\Sigma}$ とする正規分布を表す．m 番目の要素の重みは α_m で表し，混合数を M とする．EV-GMM では，S 人の事前収録話者を利用して，出力話者の平均ベクトル $\boldsymbol{\mu}_m^{(Y)}$ をバイアスベクトル $\boldsymbol{b}_m^{(0)}$ と $K(< S)$ 個の表現ベクトルの線形結合で表す．このとき，出力話者の話者性は K 次元の重みベクトル $\boldsymbol{w}^{(s)}$ で制御できる．すなわち話者空間が K 個の基底スーパーベクトル $\boldsymbol{B} = [\boldsymbol{B}_1^\top, \boldsymbol{B}_2^\top, \ldots, \boldsymbol{B}_m^\top]^\top \in \mathcal{R}^{2DM \times K}$ とバイアススーパーベクトル $\boldsymbol{b} = \left[\boldsymbol{b}_1^{(0)^\top}, \boldsymbol{b}_2^{(0)^\top}, \ldots, \boldsymbol{b}_m^{(0)^\top}\right]^\top \in \mathcal{R}^{2DM \times 1}$ によって構築される．

2.2 EVC における声質空間構築

主成分分析（PCA）に基づいて，EV-GMM を構築する場合，最初に出力話者非依存の GMM（TI-GMM）を，全ての参照話者と事前収録話者とのパラレルデータを用いて学習する．次に，対応するパラレルデータを用いて TI-GMM の出力話者の平均ベクトルを更新することで，話者依存のモデルを得る．話者空間の特徴量ベクトルとして，事前収録話者の GMM の各要素の平均ベクトルを連結し，スーパーベクトルを生成する．得られたスーパーベクトルを用いて PCA を行うことで，バイアスベクトル $\boldsymbol{b}$ と表現ベクトル $\boldsymbol{B}$ を得ることができる．このときバイアスベクトル $\boldsymbol{b}$ は声質空間における原点，表現ベクトル $\boldsymbol{B}$ は声質空間の基底と解釈できる．特定の声質はこの空間における一点で表すことができ，声質空間上での操作が出力される声質に影響する．

2.3 テンソル表現任意話者変換における声質空間

テンソル表現に基づく任意話者声質変換 (TASC) は，EVC における声質空間構築を改善した手法である [6]．EVC では前述の通り，複数のパラレルデータより得られた結合確率密度分布から，事前収録話者の話者 GMM をそれぞれ抽出し，ガウス分布の平均ベクトルを連結した GMM スーパーベクトルを用いて話者空間を構築する．しかし，GMM スーパーベクトルによる話者空間表現は，複数要因からの音響的な変動を一つの特徴量空間に含んでいる．すなわち，GMM のガウス分布の要素と平均ベクトルの次元が混在した高次の特徴量空間となっている．TASC では，任意の話者はスーパーベクトルではなく，行および列がそれぞれ GMM の要素と平均ベクトルに対応するような行列の形で表現される．このような話者表現を用いることで事前収録話者のデータセットが 3 階のテンソルで表現でき，テンソル解析を導入することで話者空間を構築することができる．テンソル解析は複数要因からの変動を適切に扱うことが可能であり [10]，より精緻な話者空間を構築できる．一対多 TASC における任意話者の平均ベクトル群 $\boldsymbol{\mu}^{(new)}$ は以下のような行列で表現される．

$$\boldsymbol{\mu}^{(new)} = \boldsymbol{U}^{(M)} \boldsymbol{W}_{(new)}^\top + \boldsymbol{b}' \quad (3)$$

ここで，$\boldsymbol{b}' = \left[\boldsymbol{b}_1^{(0)}, \boldsymbol{b}_2^{(0)}, \ldots, \boldsymbol{b}_m^{(0)}\right]^\top$ はバイアス行列であり，声質空間の原点である．$\boldsymbol{U}^{(M)} \in \mathcal{R}^{M \times K} (K \leq M)$, $\boldsymbol{W}_{(new)} \in \mathcal{R}^{D' \times K}$ はそれぞれ，表現行列および重み行列である．すなわち TASC によって構築された声質空間では，一つの声質は $D' \times K$ の行列の形で表現される．表現行列は，行列代数における特異値分解を拡張した Tucker 分解を用いることで得ることができる [6], [11]．$\boldsymbol{U}^{(M)}$ は GMM における異なる要素分布の関係性を記述していると考えられ，$\boldsymbol{W}$ を推定することで，M 混合の GMM を効率的に推定していると解釈できる．

2.4 教師なし適応を用いた声質の推定

任意の話者に対する前述の重みベクトル及び重み行列は，出力話者の音声データを用いて，最尤基準に基づいて重み $\boldsymbol{w}$（TASC では $\boldsymbol{W}$）を更新することで推定できる [5]．今，EVC の場合に，出力話者の音響特徴量系列を $\boldsymbol{Y}^{(tar)}$ とすると，$\boldsymbol{w}$ は以下のように推定できる．

$$\begin{aligned} \hat{\boldsymbol{w}} &= \underset{\boldsymbol{w}}{\operatorname{argmax}} \int P(\boldsymbol{X}, \boldsymbol{Y}^{(tar)}|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) d\boldsymbol{X} \\ &= \underset{\boldsymbol{w}}{\operatorname{argmax}} P(\boldsymbol{Y}^{(tar)}|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) \end{aligned} \quad (4)$$

ここで，出力の確率密度分布は GMM で表される．よって以下の補助関数を導入し，EM アルゴリズムを用いることで，重みを繰り返し最適化していく．

$$Q(\boldsymbol{w}, \hat{\boldsymbol{w}}) = \sum_{\boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{Y}^{(tar)}, \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) \log P(\boldsymbol{Y}^{(tar)}, \boldsymbol{m}|\boldsymbol{\lambda}^{(EV)}, \hat{\boldsymbol{w}}) \quad (5)$$

式 (5) から，$\hat{\boldsymbol{w}}$ に関する以下の更新式を得る．

$$\hat{\boldsymbol{w}} = \left\{ \sum_{m=1}^{M} \overline{\gamma}_m^{(tar)} \boldsymbol{B}_m^\top \boldsymbol{\Sigma}_m^{(YY)^{-1}} \boldsymbol{B}_m \right\}^{-1} \sum_{m=1}^{M} \boldsymbol{B}_m^\top \boldsymbol{\Sigma}_m^{(YY)^{-1}} \overline{\boldsymbol{Y}}_m^{(tar)} \quad (6)$$

$$\overline{\gamma}_m^{(tar)} = \sum_{t=1}^{T} \gamma_{m,t}, \quad \overline{\boldsymbol{Y}}_m^{(tar)} = \sum_{t=1}^{T} \gamma_{m,t} (\boldsymbol{Y}_t^{(tar)} - \boldsymbol{b}_m^{(0)}) \quad (7)$$

$$\gamma_{m,t} = P(m|\boldsymbol{Y}_t^{(tar)}, \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) \quad (8)$$

式 (6) はおよそ話者空間の基底ベクトルへの射影重みを推定していることに相当する．なお，式 (8) の初期化に際しては，TI-GMM を用いる．適応後のパラメータ生成については [12] と同様である．TASC の場合も，式 (4) に基づいてパラメータを更新することで声質を表現する重み行列 $\boldsymbol{W}$ を得ることができる [6].

任意話者声質変換における重みパラメータの推定は，出力話者の発話内容を知る必要がないため，完全な教師なし適応となる．これにより話声および歌声から声質を推定する場合，読み上げ内容および歌詞の内容を問わず，声質に関連する特徴を記述できると考えられる．すなわち，用いる音声データに含まれる言語的情報の影響を除去していると捉えることができる．また EVC 及び TASC ともに，推定パラメータ数が少ないため，極少量のデータを用いるだけで声質を適切に推定することが可能である．

3. 声質空間上での変換による声色転写

本章では，前章で述べた声質空間を用いて，「歌声らしさ」を任意の話者に転写する手法について述べる．「歌声らしさ」を適切に定義することは容易ではない．本研究では，以下「同一話者における話声と歌声の違い」に歌声らしさが表出すると考える．話声に歌声らしさを転写することは話声に対して上述の違いを表す変換を適用することに相当する．以下ではこれを声色転写と呼ぶ．本研究では，この変換をテンソル表現に基づく任意話者変換を用いて構築した声質空間上で実現した．

今，話者 A の話声の音声データ $\boldsymbol{Y}_A^{(spk)}$ が得られている条件のもと，話者 B の話声データ $\boldsymbol{Y}_B^{(spk)}$ および歌声データ $\boldsymbol{Y}_B^{(sing)}$ によって話者 B の歌声らしさを話者 A の話声に転写する場合を考える．提案する声色転写法は以下のようなプロセスで行う．

(1) 話者 A の話声データ $\boldsymbol{Y}_A^{(spk)}$ によって，この声質を表現する重み行列 $\boldsymbol{W}_A^{(spk)}$ を推定する．
(2) 話者 B についても，話声データ $\boldsymbol{Y}_B^{(spk)}$ および歌声データ $\boldsymbol{Y}_B^{(sing)}$ を用いて，それぞれの声質に対応する重み行列 $\boldsymbol{W}_B^{(spk)}, \boldsymbol{W}_B^{(sing)}$ を推定する．
(3) 話者 B において，話声を表す重み $\boldsymbol{W}_B^{(spk)}$ から $\boldsymbol{W}_B^{(sing)}$ へと変換する写像 f を推定する．すなわち，$\boldsymbol{W}_B^{(sing)} = f(\boldsymbol{W}_B^{(spk)})$ となる．
(4) 前項で求めた写像 f を用いて，$\hat{\boldsymbol{W}}_A = f(\boldsymbol{W}_A^{(spk)})$ とする．

このとき $\hat{\boldsymbol{W}}_A$ は，**話者 B の歌声らしさ を話者 A の話声に転写した声質** を表現していると考えられる．

提案法では $\boldsymbol{W}_A^{(spk)}, \boldsymbol{W}_B^{(spk)}, \boldsymbol{W}_B^{(sing)}$ を推定する際に，それぞれの話者の任意の音声データを用いることができる．特に，話者 B の歌声らしさを推定する際に話者 B の話声 $\boldsymbol{Y}_B^{(spk)}$ および歌声 $\boldsymbol{Y}_B^{(sing)}$ をパラレルデータとする必要がないのは，本手法の利点の一つである．また (3) における変換写像 f として，例えば以下のようなものが考えられる．

- 行列差分
  $\boldsymbol{W}_B^{(sing)} = \boldsymbol{W}_B^{(spk)} + \Delta\boldsymbol{W}_B$
- 線形変換
  $\boldsymbol{W}_B^{(sing)} = \boldsymbol{A}\boldsymbol{W}_B^{(spk)}$

行列差分を用いた場合は，話声と歌声の声色の差分を転写していると解釈可能である．なお声質空間を EVC に基づいて重みベクトルで構築し，写像を差分としたアルゴリズムとして，固有声に基づくキャラクター変換がある [13]. なお本研究では，歌声らしさの転写の初期検討として，行列差分を変換写像として用いた．

4. 声色転写の実験的評価

4.1 使用したデータ

提案する声色転写によって，歌声らしさが付与されたモデルを話声モデルから構築可能であることを確かめるため，実験を行った．本章では，その実験評価について述べる．

声色転写に用いる実験データとして，9 名の話者（男性 7 名，女性 2 名）から「春が来た」などの童謡 8 曲の歌唱音声を収録した．また話声データとして，童謡の歌詞を朗読した音声も合わせて収録した．収録は 48 kHz サンプリング・16 bit 量子化で行い，実験に用いる際にはすべて 16 kHz にダウンサンプリングしたものを用いた．

まず収録データの各話者の話声と歌声の違いを定量的に評価するため，ダウンサンプリングした音声を用いて聴取実験を行った．被験者は，同一話者の歌唱音声と朗読音声を聴取し，その 2 つがどの程度異なるかを 5 段階で評価した．数値が大きいほど違いが大きいとした．用いたサンプルは各話者 1 曲（春が来た） で，6 名の被験者はそれぞれすべての話者を評価した．この知覚的評価値をもとに，6 名の平均評価値が 4.1 以上の話者を「知覚的差異の大きい」 話者，4.1 未満の話者を「知覚的差異の小さい」話者とした．

4.2 声質空間及び変換モデルの構築

歌声を入力とする多対多声質変換実験によって，提案法の有効性について確認した．まず声質空間の構築及び多対多声質変換のための変換モデルを構築した．多対多の変換モデルの事前モデルとしての一対多モデルについて，参照話者として ATR 日本語音声データベース [14] から男性 1 名のデータを用いた．また事前収録話者として JNAS から男性話者 137 名，女性話者 136 名の計 273 名の発声を用いた [15]．各事前収録話者は 50 文を読み上げている．また GMM の混合数（M）は 128 とした．Tucker 分解によっ

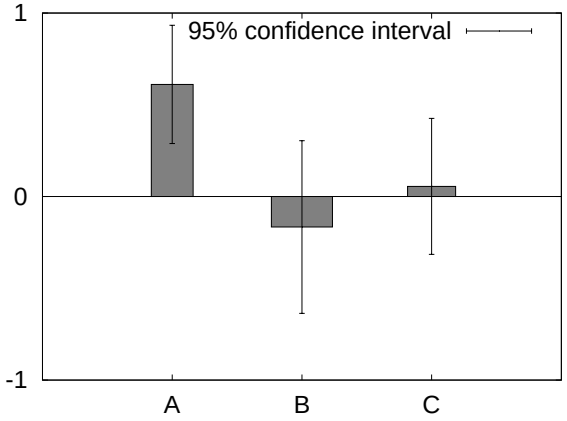

図 1 聴取実験の結果; +1 が差分転写を行った方が「歌声らしい」場合，-1 が差分転写を行わない方が「歌声らしい」場合

て表現行列を導出した後，変換モデルを精緻化するため話者正規化学習を行った [16]．このように構築した重み行列の特徴量空間を声質空間とした．行列サイズは $D' \times K$ で $D' = 48, K = 40$ である．なお多対多の変換モデルは，上記のように構築した変換モデルを分布共有させたものを用いた [17]．

スペクトル特徴量として，STRAIGHT 分析に基づくスペクトルから得られた 24 次のメルケプストラムとその動的特徴量を用いた [18]．STRAIGHT による合成に用いる非周期性指標については全周波数において-30 dB とした．

4.3 声色転写実験

歌声データを用いた声質変換を行う際，声色転写の有無によって変換音声の品質がどのように変化するかを聴取実験によって確認した．まず各話者の声質を表す重み行列について話声，歌声それぞれ，8 曲の音声データを用いて推定した．推定された重み行列を用いて，それぞれの話者について歌声らしさを表す差分行列 $\Delta \boldsymbol{W}$ を得た．得られたそれぞれの話者の差分行列を，別の話者の話声を表す声質行列 $\boldsymbol{W}^{(spk)}$ に合わせることで，変換対象となる歌声モデルを構築した．また比較対象として $\boldsymbol{W}^{(spk)}$ による歌声モデルも構築した．これら 2 つを比較することで歌声らしさの転写が行われているかを確認した．得られた歌声モデルを出力話者のモデルとし，歌唱音声を入力とした多対多声質変換を行い，歌唱合成音声を得た．

聴取実験には 6 名の被験者が参加し，各被験者は，差分の有無の異なる 8 対の合成音声について，2 つの音声 A, B を順に聴取し，それぞれの音声のうち「A の方が歌声らしい」「B の方が歌声らしい」「同じ」の 3 択で評価した．8 対の合成音声を知覚的差異の大きさ及び差分行列のフロベニウスノルムの大きさによって以下の 3 つのグループに区分した．

A 知覚的差異が小さくフロベニウスノルムが大きい (3)
B 知覚的差異が大きくフロベニウスノルムが小さい (2)
C 知覚的差異が小さくフロベニウスノルムが小さい (3)
括弧内はペア数を表す．

聴取実験の結果を図 1 に示す．図 1 より，差分行列のフロベニウスノルムが大きい場合（A） は，声色転写を行った場合に歌声らしさが向上することが分かる．一方，話声と歌声の知覚的差異が大きい話者についても，差分行列のフロベニウスノルムが小さい場合（B） は差分転写による有効性は現れなかった．このような話者は，歌声らしさが声道特性に起因する特徴とは異なる特徴によって表出していると考えられ，$F0$ パターンの変換も含めた歌声らしさの転写が必要になると考えられる．

以上より声質空間上での写像を別の話者に適用する提案手法は，多様な歌声モデルを構築する上で一定の効果があると考えられる．

5. おわりに

本稿では，特定話者の声に存在する歌声らしさを抽出し，話声データによって構築された話者モデルに付加することで，歌声モデルを構築する手法を提案した．提案手法では，任意話者声質変換で用いられる重みベクトルや重み行列に着目し，この特徴量空間を声質を定量的に表す声質空間であると考える．同一話者内での話声と歌声の違いを声質空間上での変換として捉え，この変換を異なる話者の話声特徴に対して適用することで，「話声と歌声の違い」の転写を実現する．多対多歌声声質変換の実験によって，重み行列の差分が大きい場合には適切に歌声らしさが転写されることを示した．

今後の課題として，VocaListener2 が取り扱っているような声色の時間的変化を提案法の枠組みでどのようにモデル化するかが上げられる．加えて，今回の手法では歌声らしい音響特徴に関する知見を導入していないため，これを声質空間上で適切に表現することで品質向上が期待できる．また今回は変換写像としてもっとも単純な差分を用いたが，声質空間上での線形変換を推定することで，どのような転写が実現可能かについても検討していく予定である．

謝辞 本研究の一部は科研費・研究活動スタート支援 (23800015) の助成を受けたものである．

参考文献

[1] 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.
[2] 中野倫靖, 後藤真孝, “VocaListener2: ユーザ歌唱の音高と音量だけでなく声色変化も真似る歌声合成システムの提案,” 情報処理学会研究報告 音楽情報科学 2010-MUS-86, no. 3, pp.1–10, 2010.
[3] 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, no. 5, pp. 25–32, 2008.
[4] 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, ” 情報処理学会研究報告 音楽情報科学 2010-MUS-86, no. 1, pp. 1–8, 2010.
[5] T. Toda, Y. Ohtani, and K. Shikano, “Eigenvoice conversion based on Gaussian mixture model,” Proc. INTERSPEECH, pp. 2446–2449, 2006.
[6] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, “One-to-many voice conversion based on tensor representation of speaker space,” Proc. INTERSPEECH, pp. 653–656, 2011.
[7] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” Proc. ICASSP, vol. 1, pp. 285–288, 1998.
[8] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, “High-performance robust speech recognition using stereo training data,” Proc. ICASSP, pp. 301–304, 2001.
[9] A. Kunikoshi, Y. Qiao, N. Minematsu, and K. Hirose, “Speech generation from hand gestures based on space mapping,” Proc. INTERSPEECH, pp. 308–311, 2009.
[10] M. A. O. Vasilescu and D. Terzopoulos, “Mutilinear analysis of image ensembles: TensorFaces,” Proc. ECCV, pp. 447–460, 2002.
[11] L. R. Tucker, “Some mathematical notes on three-mode factor analysis,” Psychometrika, vol. 31, no. 3, pp. 279–311, 1966.
[12] T. Toda, A. W. Black, and K. Tokuda, “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.
[13] T. Pongkittiphan, D. Saito, N. Minematsu, K. Hirose, “Eignvoice-based character conversion and its evaluation,” IEICE Technical Report, SP2012-34, pp. 7–12, 2012.
[14] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K.Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol.9, pp.357–363, 1990.
[15] “Jnas: Japanese newspaper article sentences,” http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html
[16] D. Saito, N. Minematsu, and K. Hirose, “Effects of speaker adaptive training on tensor-based arbitrary speaker conversion,” Proc. INTERSPEECH2012 (to appear).
[17] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Non-parallel training for many-to-many eigenvoice conversion,” Proc. ICASSP, pp. 4822–4825, 2010.
[18] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol.27, pp.187–207, 1999.