一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2014-30(2014-05)

統計的手法に基づく歌声の知覚年齢制御法

小林　和弘† 　戸田　智基† 　中野　倫靖†† 　後藤　真孝†† 　Graham Neubig†

Sakriani Sakti† 　中村　哲†

† 奈良先端科学技術大学院大学 情報科学研究科
〒 630–0192 奈良県生駒市高山町 8916 番地の 5
†† 独立行政法人 産業技術総合研究所 情報技術研究部門
〒 305-8568 茨城県つくば市梅園 1-1-1
E-mail: †{kazuhiro-k,tomoki,neubig,ssakti,s-nakamura}@is.naist.jp, ††{t.nakano,m.goto}@aist.go.jp

あらまし　歌声に対する主観的情報である知覚年齢は，歌声の特徴を直感的に記述できる要素の一つである．歌手は，音高や声質を変化させることで様々な歌声を生み出すことができる．しかし，声質は歌手の身体的特徴に制限されており，身体的制約を超えた声質での歌唱は困難である．これに対して，身体的制約を超える歌唱を実現する手法として，統計的手法に基づく歌声声質変換が提案されている．統計的手法に基づく歌声声質変換は，音響特徴量の対応関係を統計的にモデル化することで，入力歌手の声質を目標歌手の声質へと変換することが可能である．ただし，変換後の声質は目標歌手のものに限定されるため，性別や年齢のような直感的に理解しやすい基準に沿って変換歌声の声質を自由に制御する事は困難である．本報告では，入力歌手の個人性を保持しつつ，知覚年齢を操作することで声質を制御する技術として，我々が提案した統計的手法に基づく歌声声質変換について述べる．実験結果より，提案法は歌手の個人性に悪影響を与えずに，知覚年齢に沿った声質制御が可能である事を示す．
キーワード　歌声声質変換，知覚年齢，分節的特徴，個人性，声質制御

Statistical approach to perceived age control of singing voice

Kazuhiro KOBAYASHI†, Tomoki TODA†, Tomoyasu NAKANO††, Masataka GOTO††, Graham NEUBIG†, Sakriani SAKTI†, and Satoshi NAKAMURA†

† Information Science, Nara Institute of Science and Technology (NAIST)
Takayama 8916–5, Ikoma, Nara, 630–0192 Japan
†† Information Technology Research Institute,
National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568 Japan
E-mail: †{kazuhiro-k,tomoki,neubig,ssakti,s-nakamura}@is.naist.jp, ††{t.nakano,m.goto}@aist.go.jp

Abstract　The perceived age of a singing voice is the age of the singer as perceived by the listener, and is one of the notable characteristics that determines perceptions of a song. Singers can sing expressively by controlling prosody and voice timbre, but the varieties of voice timbre that singers can produce are limited by physical constraints. Previous work has attempted to overcome the limitation through the use of statistical singing voice conversion. This technique makes it possible to convert singing voice timbre of an arbitrary source singer into those of an arbitrary target singer. However, it is still difficult to intuitively control singing voice characteristics by manipulating parameters corresponding to specific physical traits, such as gender and age. This paper presents our recent work on the development of a technique for controlling the voice timbre based on perceived age while maintaining the singer's individuality. The experimental results show that the proposed voice timbre control method makes it possible to change the singer's perceived age while not having an adverse effect on the perceived individuality.
Key words　singing voice conversion, perceived age, segmental features, individuality, voice quality control.

1. はじめに

歌声は，メロディーやリズムに加え言語情報である歌詞を伝えることができ，他の楽器には難しいような多彩な音楽表現が可能である．更に歌手の技量によるものの，声質に関しても声帯や調音器官を巧みに操ることで変化させる事が可能である．しかしながら，個々の歌手が表現できる声質の範囲は身体的制約により制限されている．この身体的制約を超え，歌手が意のままに操れる自在な声質制御が実現すれば，更なる豊かな音楽表現を生み出すことが期待される．

歌手の声質を制御する手法として様々な手法が提案されている．代表的な手法として，音声分析合成処理によるモーフィング [1] がある．モーフィングは，異なる歌手が歌う同一楽曲間において，スペクトル特徴量や基本周波数（F_0）を各々独立に補間することで，新たな歌声を合成する事が可能である．しかし，モーフィングによる声質制御は同一楽曲間のみを対象とするため利用範囲は限定される．

任意の楽曲において声質を変換する技術として，統計的手法に基づく声質変換 [2], [3] を歌声へと適用した統計的手法に基づく歌声声質変換（SVC: Singing Voice Conversion）[4], [5] が提案されている．SVC は，学習処理と変換処理に分かれており，学習処理では入力歌手と目標歌手の同一曲の歌声データ（パラレルデータ）に基いて，両歌手の音響特徴量の対応関係を混合正規分布モデル（GMM: Gaussian Mixture Model）により学習する．変換処理では，学習済み GMM に基づき，任意の楽曲に対して入力歌手の歌声を目標歌手の声質を持つ歌声へと変換する．さらに，任意の入力歌手と任意の目標歌手の間での声質変換を可能とする固有声変換技術 [6], [7] を歌声に適用した手法が提案されている [8]．多対多固有声に基づく SVC [8] では，一人の参照歌手と多数の事前収録目標歌手のパラレルデータセットを用いて，固有声 GMM（EV-GMM: Eigenvoice GMM）の学習を行う．任意の入力歌手および目標歌手に対する変換モデルは，各歌手の極少量の歌声データを用いて，EV-GMM の適応パラメータを各々独立に推定することで，容易に構築することが可能である．また，EV-GMM の適応パラメータを手動操作することで，変換歌声の声質を変化させることも可能である．しかし，適応パラメータ空間は人の主観軸に沿わないため，パラメータ操作後の変換歌声の声質は予測困難であり，直感的な声質制御は困難である．歌手の思い描く理想の声質へと変換可能な声質制御技術を実現するためには，直感的な声質制御の実現が必要である．

統計的パラメトリック音声合成の研究において，性別や年齢などの直感的な声質制御パラメータにより手動で声質を制御する手法が提案されている．隠れマルコフモデル（HMM: Hidden Markov Model）に基づくテキスト音声合成 [9] において，発話様式を表す低次元ベクトルから HMM の平均ベクトルへの写象を表現する重回帰 HMM を用いることで，合成音声の発話様式を手動制御することが可能となる [10]．さらに，“暖かい–冷たい” などの声質表現語対に対する主観評価値で構成される低次元ベクトルを導入することで，合成音声の声質を直感的に制御することが可能となる [11]．類似した枠組みとして，韻律パラメータと感情を表すパラメータに対する重回帰分析に基づき，感情音声を合成する手法も提案されている [12]．なお，重回帰 HMM は歌声合成にも適用されており，学習に用いた単一歌手内での “子供らしい–大人びた” といった主観的制御パラメータに基づいた歌唱様式の操作法が提案されている [13]．

歌声の直感的な声質制御を実現する上で，話声における声質表現語対 [14] のように声質を主観的に表す尺度がいくつか考えられるが，本報告では歌声の知覚年齢 [15] に着目する．知覚年齢に沿った声質制御を実現することで，万人が持つ年齢という基準により声質を制御可能となる．

本報告では，歌手の個人性を保持したまま歌声の知覚年齢を制御する技術として，我々が提案した統計的手法に基づく歌声声質制御法について述べる．まず，歌声の知覚年齢に寄与する音響特徴量を明らかにするために，F_0 や継続長を表す韻律的特徴とスペクトル特徴量や非周期成分を表す分節的特徴に対し，どちらの特徴が歌声の知覚年齢及び歌手の個人性に大きく寄与しているかを調査した結果について述べる．そして，歌手の個人性を保持しつつ知覚年齢に基づき声質を制御する手法について述べ，その有効性を示す．

2. 統計的手法に基づく歌声声質変換

GMM に基づく SVC は，歌手の声質を異なる歌手の声質へと変換する技術である．GMM に基づく SVC は学習処理と変換処理で構成される．

学習処理では，話声における声質変換と同様に，入力歌手と目標歌手のパラレルデータより音響特徴量を抽出し，GMM によりそれらの結合確率密度関数をモデル化する．入力歌手と目標歌手の音響特徴量を，各々$2D$ 次元の静的動的特徴量ベクトル $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$，$\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ とする．ここで，$\boldsymbol{x}_t$ と $\boldsymbol{y}_t$ は，フレーム t における入力歌手と目標歌手の静的音響特徴量であり，$\Delta\boldsymbol{x}_t$ と $\Delta\boldsymbol{y}_t$ は，同フレームの入力歌手と目標歌手の動的特徴量である．$\top$ は転置を表す．これらの音響特徴量の結合確率密度関数は，以下の式により与えられる．

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right), \quad (1)$$

ここで $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は，平均ベクトル $\boldsymbol{\mu}$，共分散行列 $\boldsymbol{\Sigma}$ の正規分布を表す．混合数は M であり，m は分布番号を表す．$\boldsymbol{\lambda}$ は GMM のパラメータセットを表し，個々の分布における分布重み α_m，平均ベクトル $\boldsymbol{\mu}_m$，共分散行列 $\boldsymbol{\Sigma}_m$ を含む．パラレルデータに対して，動的時間伸縮により対応づけられた $\boldsymbol{X}_t$，$\boldsymbol{Y}_t$ を用いて GMM を学習する．

変換処理では，入力歌手の歌声から抽出された音響特徴量を最尤推定法 [3] により目標歌手の音響特徴量へと変換する．入力歌手と目標歌手の特徴量系列ベクトルを，各々$\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ と $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$ とする．ここで，T はフレーム数である．変換された静的特徴量系列 $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ は次式で示される．

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) \text{ subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \quad (2)$$

ここで $\boldsymbol{W}$ は静的特徴量系列を結合静的動的特徴量系列に拡張する行列である．条件付き確率密度関数 $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})$ は，式 (1) で与えられた結合確率密度関数から解析的に導出される．なお，過剰な平滑化による変換歌声の音質劣化を緩和するため，系列内変動（GV: Global Variance）[3] を考慮する．

3. 知覚年齢に寄与する音響特徴量の調査

本節では，歌声を対象とした韻律的特徴と分節的特徴の知覚年齢への寄与を調査した結果について述べる．また，個人性を保持した知覚年齢に基づく声質制御を実現するために韻律的特徴と分節的特徴のどちらがより歌手の個人性を含んでいるかを調査した結果についても述べる．

3.1 韻律的特徴と分節的特徴の知覚年齢への寄与

SVC により，メルケプストラムと非周期成分を変換することで，入力歌手から目標歌手へと声質を変換した変換歌声（SVC

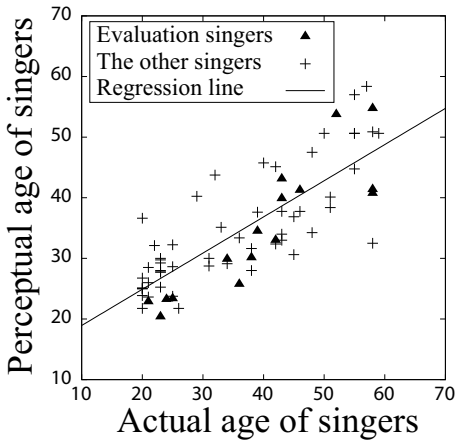

図 1　歌手の実年齢と歌声の知覚年齢スコアの相関

歌声）を合成する．得られる SVC 歌声は，入力歌手の F_0 パターン，パワーパターン，継続長といった韻律的特徴と目標歌手のメルケプストラム，非周期成分といった分節的特徴を併せ持つ．この変換歌声の知覚年齢と，同一歌手 SVC（入力）および同一歌手 SVC（目標）の知覚年齢をそれぞれ比較することで，分節的特徴と韻律的特徴のどちらの音響特徴量がより知覚年齢に寄与するかを調査する．なお，同一歌手 SVC（入力/目標）とは，入力（目標）歌手から同一の入力（目標）歌手への SVC であり，多対多固有声変換 [7] を応用することで，入力（目標）歌手から参照歌手への GMM を用いて実現することができる [15]．これにより，入力（目標）歌手の音響特徴量に対して，GMM に基づく変換によって生じる変換誤差を与えた際の歌声を合成することができる．

3.2　韻律的特徴と分節的特徴の個人性への寄与

SVC における知覚年齢に沿った声質制御では，歌手の個人性を保ったまま知覚年齢のみを操作できる手法を確立することが望まれる．そのため歌声の個人性が，韻律的特徴と分節的特徴のどちらに多く反映されているかを調査する．

3.3　音響特徴量の調査に関する実験的評価

3.3.1　実 験 条 件

評価データベースとして，20，30，40，50 歳代の日本人男女の歌唱データを含む，AIST ハミングデータベース：ポピュラー音楽（RWC-MDB-P-2001）[16] を用いる．歌手の総数は 75 名であり，各歌手における曲数は 25 曲である．各曲の長さは 20 秒程度である．評価楽曲として 1 曲の評価楽曲（P039）を用いる．

初めに聴取実験による歌手の実年齢と歌声の知覚年齢の対応関係を調査する．20 代男性 8 名の被験者が，全歌手の評価楽曲に対し知覚年齢スコアを付与する．

知覚年齢に寄与する音響特徴量の特定のため，SVC 変換歌声と同一歌手 SVC 歌声（入力/目標）の知覚年齢スコアを比較する．20 歳代男性 8 名の被験者が，各変換歌声に対し知覚年齢スコアを付与する．20 歳代から 50 歳代の各年代の実年齢を持つ男女 2 名ずつ，計 16 名を評価歌手とする．全年代かつ男女の評価歌手が割振られるように評価歌手を 2 グループに分け，各被験者は，1 グループに対して知覚年齢スコアを付与する．

次に SVC 変換歌声と同一歌手 SVC 歌声（入力/目標）を用いて，知覚年齢変換処理における個人性の変化を調査する．評価歌手全 16 名を男女を区分した全年代を網羅する 4 名ずつの 4 セットに分け，各セット内における評価歌手の総当りペアに対して SVC による変換歌声（12 種類）を作成する．被験者は，同一の歌手が歌っているという評価基準のもと，SVC による変換歌声と，各セットにおける個々の評価歌手の同一歌手 SVC 歌声（入力/目標）を比較し，どの評価歌手に最も近いか判断する．また，被験者に対し，同一の歌手においても年齢が変化しているという可能性を予め伝えて実験を行う．被験者は，各セットごとに 2 名の計 8 名である．

サンプリング周波数は 16 kHz である．音響特徴量として STRAIGHT [17] で抽出されたメルケプストラム係数の 1 次から 24 次を用いる．音源情報は，F_0 と 0–1，1–2，2–4，4–6，6–8 kHz の 5 周波数帯で平均化された非周期成分を用いる．フレームシフト長は 5 ms である．

同一歌手 SVC（入力/目標）において，メルケプストラム及び非周期成分を変換するための GMM を作成するため，参照歌手として評価歌手以外の歌手を 1 名用いる．異なる歌手間の SVC においては，各グループ内の評価歌手の総当りペアに対して，個別にメルケプストラム及び非周期成分を変換するための GMM を学習し，変換を行う．

3.3.2　実 験 結 果

図 1 に歌手の実年齢と歌声の知覚年齢の相関図を示す．横軸は歌手の実年齢であり，縦軸は各歌手に対する知覚年齢スコアの平均値である．各歌手に対する被験者の知覚年齢の標準偏差は 6.17 である．全体の相関係数は 0.81 であり，歌手の実年齢と知覚年齢に対して強い相関がみられる．

図 2，3 に，同一歌手 SVC 歌声（入力/目標）の知覚年齢スコアと SVC 歌声の知覚年齢スコアの相関を示す．図 2 は，横軸を同一歌手 SVC 歌声（入力）の知覚年齢スコアにしたものであり，韻律的特徴が知覚年齢への寄与が大きい場合，相関が高くなる．図 3 は，横軸を同一歌手 SVC 歌声（目標）の知覚年齢スコアにしたものであり，分節的特徴の知覚年齢への寄与が大きい場合，相関が高くなる．どちらの図においても，正の相関が観測されることから，韻律的特徴および分節的特徴のどちらも知覚年齢に影響を与えることが分かる．また，韻律的特徴は，分節的特徴に比べ，より大きく知覚年齢に寄与することが分かる．

表 1 に，SVC において韻律的特徴もしくは分節的特徴の変換を行った際に生じる歌手の個人性の変化に対する評価結果を示す．表は，入力歌手の韻律的特徴と目標歌手の分節的特徴を持つ SVC 歌声が，韻律的特徴が一致する同一歌手 SVC 歌声（入力）に似ていると判断された場合の確率，分節的特徴が一致する同一歌手 SVC 歌声（目標）に似ていると判断された場合の確率，入力歌手と目標歌手以外の同一歌手 SVC 歌声（その他）に似ていると判断された場合の確率をそれぞれ表す．表より，歌手の個人性は，分節的特徴に比べ韻律的特徴で識別される傾向が強いことがわかる．図 2，3 の結果と同様の傾向であることから，歌手の個人性と知覚年齢の相関は高いといえる．

4.　知覚年齢に基づく歌声声質制御

本節では，統計的手法に基づく歌声声質変換（SVC: Singing Voice Conversion）において入力歌手の個人性を保持したまま，歌手の知覚年齢を操作する手法について述べる．3.3 節の実験結果で示した通り，韻律的特徴は知覚年齢に大きく寄与しているものの，歌手の個人性にも大きく影響を与える．そこで，個人性を保持した知覚年齢制御を実現するための第一歩として，個人性に与える影響の小さい分節的特徴を操作対象とする知覚年齢制御法について述べる．まず重回帰混合正規分布モデル（MR-GMM: Multiple-regression Gaussian Mixture Model）に基づく声質変換 [18] を SVC に適用する．さらに，MR-GMM に基づく SVC に対して，多対多固有声 GMM（EV-GMM: Eigenvoice GMM）に基づく SVC の枠組みを適用し，多対多 MR-GMM に基づく SVC を実現することで任意

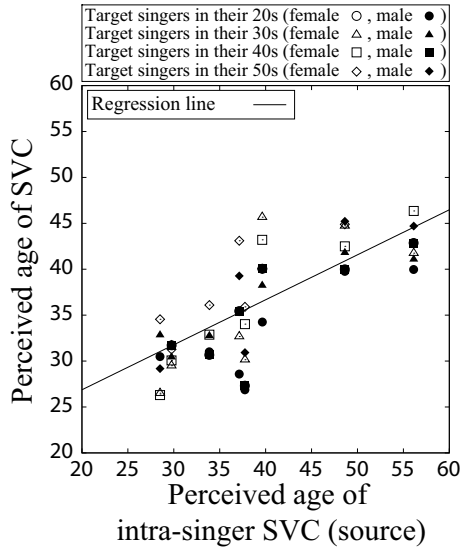

図 2　同一歌手 SVC 歌声（入力）と SVC 歌声の知覚年齢の相関

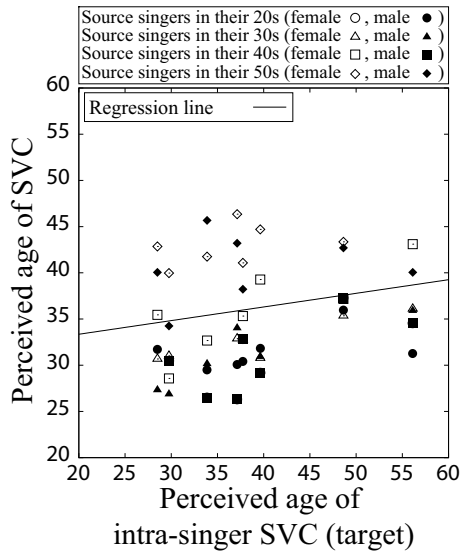

図 3　同一歌手 SVC 歌声（目標）と SVC 歌声の知覚年齢の相関

表 1　SVC における個人性の評価

特徴	割合
韻律的特徴	52.08
分節的特徴	35.42
不一致	12.50

の入力歌手への対応を容易にする．そして，多対多 MR-GMM における平均ベクトルの表現方法を変えることで，個人性を保持した知覚年齢制御を実現する．なお，提案法では分節的特徴により声質を変換するため，リアルタイム声質変換システム [19], [20] と組み合わせることが容易であり，リアルタイムかつ直感的な声質制御の実現が見込まれる．

4.1　重回帰混合正規分布モデルに基づく歌声声質変換

知覚年齢に基づく声質制御を実現するため，MR-GMM に基づく声質変換を SVC に適用する．MR-GMM に基づく SVC は，GMM に基づく SVC と同様に学習処理と変換処理により構成される．

学習処理では，一人の参照歌手と複数の事前収録目標歌手のパラレルデータセットより分析された音響特徴量に基づき，結合確率密度関数をモデル化する．MR-GMM による結合確率密度関数は以下のように与えられる．

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}^{(MR)}, w^{(s)}\right) = \sum_{m=1}^{M}\alpha_m\mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t\\ \boldsymbol{Y}_t\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}_m^{(X)}\\ \boldsymbol{\mu}_m^{(Y)}(s)\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)}\\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right). \quad (3)$$

s 番目の事前収録目標歌手の平均ベクトルは以下の式で与えられる．

$$\boldsymbol{\mu}_m^{(Y)}(s) = \boldsymbol{b}_m^{(Y)}w^{(s)} + \overline{\boldsymbol{\mu}}_m^{(Y)}, \quad (4)$$

ここで $\boldsymbol{b}_m^{(Y)}$ および $\overline{\boldsymbol{\mu}}_m^{(Y)}$ は，それぞれ代表ベクトルとバイアスベクトルを表す．$w^{(s)}$ は，s 番目の事前収録目標歌手の歌声に対して聴取実験により得られた知覚年齢スコアを表す．

変換処理において，MR-GMM における出力側の平均ベクトルは，所望の知覚年齢スコアを入力することで決定される．入力歌手の歌声は，2 節と同様に系列内変動を考慮した最尤系列推定法 [3] により変換される．

4.2　多対多重回帰混合正規分布モデルに基づく歌声声質変換

参照歌手のみではなく任意の入力歌手（ユーザ）の変換に容易に対応するため，多対多 EV-GMM に基づく SVC [8] で用いられる変換法を MR-GMM に基づく SVC に適用する．多対多 MR-GMM の結合確率密度関数は以下の式で表される．

$$\begin{aligned}&P\left(\boldsymbol{Y}_t^{(i)}, \boldsymbol{Y}_t^{(o)}|\boldsymbol{\lambda}^{(MR)}, w^{(i)}, w^{(o)}\right)\\ &= \sum_{m=1}^{M} P\left(m|\boldsymbol{\lambda}^{(MR)}\right)\int P\left(\boldsymbol{Y}_t^{(i)}|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(MR)}, w^{(i)}\right)\\ &P\left(\boldsymbol{Y}_t^{(o)}|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(MR)}, w^{(o)}\right)P\left(\boldsymbol{X}_t|m, \boldsymbol{\lambda}^{(MR)}\right)\mathrm{d}\boldsymbol{X}_t\\ &= \sum_{m=1}^{M}\alpha_m\mathcal{N}\left(\begin{bmatrix}\boldsymbol{Y}_t^{(i)}\\ \boldsymbol{Y}_t^{(o)}\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}_m^{(Y)}(i)\\ \boldsymbol{\mu}_m^{(Y)}(o)\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_m^{(YY)} & \boldsymbol{\Sigma}_m^{(YXY)}\\ \boldsymbol{\Sigma}_m^{(YXY)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right) \quad (5)\end{aligned}$$

$$\boldsymbol{\Sigma}_m^{(YXY)} = \boldsymbol{\Sigma}_m^{(YX)}\boldsymbol{\Sigma}_m^{(XX)^{-1}}\boldsymbol{\Sigma}_m^{(XY)}, \quad (6)$$

ここで $w^{(i)}$ と $w^{(o)}$ は，それぞれ入力歌手と目標歌手の知覚年齢スコアを示す．入力歌手と目標歌手の多対多 MR-GMM における平均ベクトルは，式（4）により決定される．

本報告では，入力歌手と参照歌手のパラレルデータが入手可能な状況を想定する．この場合においても，式（5）の入力平均ベクトル $\boldsymbol{\mu}_m^{(Y)}(i)$ を式（4）により表すことが可能である．しかし，入力平均ベクトルが代表ベクトルで張られる部分空間で表現されるために，入力歌手に対するモデル化精度が下がる事が考えられる．そこで，入力歌手と参照歌手のパラレルデータにより，最尤推定基準により多対多 MR-GMM の入力平均ベクトルを更新する．更新された入力平均ベクトルは以下の式で与えられる．

$$\boldsymbol{\mu}_m^{(Y)}(i) = \hat{\boldsymbol{\mu}}_m^{(Y)}, \quad (7)$$

ここで $\hat{\boldsymbol{\mu}}_m^{(Y)}$ はパラレルデータによって得られた入力歌手の平均ベクトルの最尤推定値である．なお，入力歌手と参照歌手のパラレルデータが入手可能であるならば，個々の入力歌手に対し，入力歌手と全事前収録目標歌手の MR-GMM を学習する事も可能である．しかしながら，MR-GMM の学習には多大な時間を要し，入力歌手の増加につれて MR-GMM の学習数も増えるため，効率が悪い．これに対して，多対多 SVC の枠組みを用いることで，式（4.2）で示される入力平均ベクトルの最尤推定のみで各入力歌手に対する高精度な MR-GMM を学習することが可能となる．

表 2　MR-GMM と Modified MR-GMM における入出力の関係

手法	入力	出力	出力平均ベクトル
MR-GMM	知覚年齢スコア $w^{(o)}$	$w^{(o)}$ 歳の平均的な声質	$\boldsymbol{b}_m^{(Y)} w^{(o)} + \overline{\boldsymbol{\mu}}_m^{(Y)}$
Modified MR-GMM	差分知覚年齢スコア Δw	$w^{(i)} + \Delta w$ 歳の声質	$\hat{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{b}_m^{(Y)} \Delta w$

4.3　個人性を保持した知覚年齢に基づく声質制御法

多対多 MR-GMM に基づく SVC により，入力歌手の声質を知覚年齢に基づいた目標歌手の声質へと変換することが可能となる．しかし，式 (4) で表現される出力平均ベクトルは，所望の知覚年齢スコアを持つ事前収録目標歌手の平均的な声質を表しており，特定歌手の声質は精度良く表現できない．入力歌手の個人性を保持しつつ知覚年齢のみを制御する声質制御法を実現するために，多対多 MR-GMM における出力平均ベクトルの表現方法を変更する．

$$
\begin{aligned}
\boldsymbol{\mu}_m^{(Y)}(s) &= \boldsymbol{b}_m^{(Y)} w^{(s)} + \overline{\boldsymbol{\mu}}_m^{(Y)} \\
&= \boldsymbol{b}_m^{(Y)} (w^{(i)} + \Delta w) + \overline{\boldsymbol{\mu}}_m^{(Y)} \\
&= \boldsymbol{b}_m^{(Y)} w^{(i)} + \overline{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{b}_m^{(Y)} \Delta w \\
&\simeq \hat{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{b}_m^{(Y)} \Delta w, \qquad (8)
\end{aligned}
$$

ここで出力側の知覚年齢スコア Δw は，入力歌手の知覚年齢を基に，そこより変化させる差分知覚年齢スコアで表される．出力平均ベクトルは，入力歌手の知覚年齢を持つ平均ベクトル $\hat{\boldsymbol{\mu}}_m^{(Y)}$ に加え，重回帰分析より得られた代表ベクトル $\boldsymbol{b}_m^{(Y)}$ と差分知覚年齢スコア Δw で決定される平均ベクトルに対する変形量を加えることで表される．入力歌手の特定モデルを基に，年齢に伴う変形量が付与されるため，歌手の個人性を保持した知覚年齢制御が期待される．

5.　知覚年齢制御の実験的評価

本評価では，4.1 節の変換法を MR-GMM，4.3 節の変換法を Modified MR-GMM を表記する．表 2 に，MR-GMM と Modified MR-GMM の入出力の関係を示す．

5.1　実 験 条 件

学習データとして，参照歌手 1 名と事前収録目標歌手 54 名（男性 27 名，女性 27 名）を用いる．参照歌手と事前収録目標歌手のパラレルデータは動的時間伸縮により予め対応付けておく．評価歌手として，事前収録目標歌手に含まれない 16 名の歌手を用いる．事前収録目標歌手の知覚年齢スコアは，20 代男性被験者 1 名により全 25 曲に付与された知覚年齢スコアの平均値を用いる．評価楽曲として，学習データに含まれる 1 曲（P039）を用いる．MR-GMM および Modified MR-GMM の混合数は，スペクトル特徴量に対しては 128，非周期成分に対しては 32 とする．

Modified MR-GMM による知覚年齢の変動精度を評価する．20 代男性の被験者 8 名と評価歌手 16 名を，各年代の評価歌手を含むように被験者 4 名と評価歌手 8 名で構成される 2 グループに分ける．式 (8) における差分知覚年齢スコアの設定は，-60, -40, -20, 0, 20, 40, 60 と変化させて変換歌声を作成する．ここで差分知覚年齢スコアが 0 の時の変換歌声は，声質操作を行わない同一歌手 SVC 歌声（入力）と等価である．被験者は，ランダムな順番で提示される変換歌声に対してそれぞれ知覚年齢の評価を行う．

XAB テストにより Modified MR-GMM と MR-GMM の個人性に関するの対比較実験を行う．被験者および評価歌手は，前実験と同様に 2 グループに分ける．式 (8) における差分知覚年齢スコアの設定は，-60, -30, 30, 60 と変化させて変換歌声を作成する．式 (4) における MR-GMM の知覚年齢スコアの設定は，前実験で使用した同一歌手 SVC 歌声（入力）に対する知覚年齢評価結果の平均値を基準に ± 30, 60 と変化させた値を用いる．各評価歌手の同一歌手 SVC 歌声（入力）を先に提示し，MR-GMM と Modified MR-GMM による変換歌声をランダムな順で再生し，どちらが同一歌手 SVC 歌声（入力）に似ているかという基準で評価を行う．その際に，被験者には歌手の知覚年齢は変化しているという事を伝える．

5 段階の平均オピニオン評点（MOS: Mean opinion score）により，Modified MR-GMM と MR-GMM に対して変換歌声の自然性の評価を行う．被験者及び評価歌手は，前実験と同様に 2 グループに分けて評価を行う．差分知覚年齢スコアおよび知覚年齢スコアの設定は前実験と同様である．被験者に対して，自然歌声および Modified MR-GMM，MR-GMM の変換歌声をランダムな順番で提示する．被験者は，それぞれの変換歌声に対して “5–とても良い”，“4 –良い”，“3–ふつう”，“2 –悪い”，“1–とても悪い” という 5 段階評価で自然性の評価を行う．

その他の実験条件は，3.3.1 節と同様である．

5.2　実 験 結 果

図 4 に，Modified MR-GMM による知覚年齢変動の評価結果を示す．横軸は差分知覚年齢スコアの設定値であり，縦軸は同一歌手 SVC 歌声（入力）の知覚年齢評価結果と各差分知覚年齢スコアの知覚年齢評価結果の差を表す．差分知覚年齢スコアの設定を，-60 から 60 にかけて変化させることで，線形性を保って知覚年齢が高くなっていく事がわかる．また，これは，3.3.2 節の図 3 における分節的特徴の寄与とほぼ同様の傾向で変化している事が確認出来る．このことより，Modified MR-GMM は分節的特徴の知覚年齢への寄与に基づいた変化を適切にモデル化出来ていることがわかる．

図 5 に，Modified MR-GMM と MR-GMM の歌手の個人性に関する対比較実験の評価結果を示す．全ての差分知覚年齢スコアの設定において，Modified MR-GMM の方が歌手の個人性をより正確に捉えた変換歌声を合成できていることから，Modified MR-GMM の有効性が確認される．

図 6 に，変換歌声の自然性に関する 5 段階評価 MOS の評価結果を示す．全ての差分知覚年齢スコアの設定値において，Modified MR-GMM の有効性が確認される．式 (8) における Modified MR-GMM のバイアスベクトル $\hat{\boldsymbol{\mu}}_m^{(Y)}$ は，入力歌手の声質を表現している．一方で，式 (8) における MR-GMM のバイアスベクトル $\overline{\boldsymbol{\mu}}_m^{(Y)}$ は，S 人の事前収録目標歌手の平均的な声質を表す．そのため，Modified MR-GMM の平均ベクトルに比べ，MR-GMM の平均ベクトルでは，変換特徴量の過剰な平滑化が起こり，変換歌声の自然性が大きく損なわれると考えられる．

図 4，5，6 より，Modified MR-GMM は，歌手の個人性を保ちつつ知覚年齢に基づく声質制御を可能とし，MR-GMM に比べ高い自然性での変換歌声が合成可能であることがわかる．

6.　結　　論

本報告では，知覚年齢に基づいた声質制御を実現するために，知覚年齢に寄与する音響特徴量の調査および歌手の個人性を保持した知覚年齢制御について述べた．音響特徴量の調査結果より，韻律的特徴は，歌声の知覚年齢に大きく寄与しているものの，歌手の個人性にも大きく寄与している事がわかった．一方で，分節的特徴は，韻律的特徴に比べ，知覚年齢への寄与は小さいものの，個人性に対する影響も小さいため，個人性を

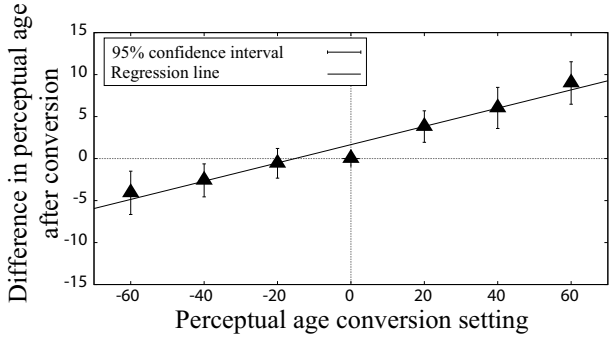

図 4　Modified MR-GMM における指定した差分知覚年齢スコアと変換歌声の知覚年齢

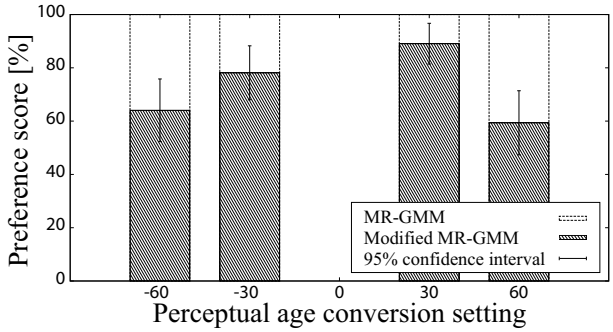

図 5　個人性に関する対比較実験結果.

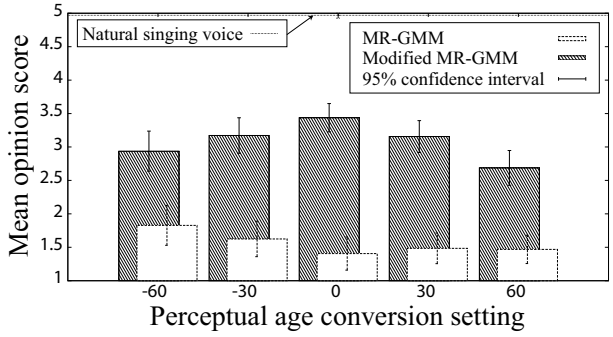

図 6　変換歌声の自然性に関する MOS 評価

保持した知覚年齢制御には，適している事がわかった．従来の重回帰混合正規分布モデル（MR-GMM: Multiple-regression Gaussian Mixture Model）に基づく声質変換では，知覚年齢に基いた声質制御は可能であるものの，変換歌声は入力した知覚年齢に基づく事前収録歌手の平均的な声質へと変換されるため，歌手の個人性を保持した声質制御は困難であった．この問題を解決するために，MR-GMM における出力平均ベクトルの表現方法を変更することで個人性を保持しつつ知覚年齢制御を可能とする手法（Modified MR-GMM）を提案した．評価結果より，Modified MR-GMM は歌手の個人性を保持しつつ知覚年齢の制御を可能とし，MR-GMM に比べ高い自然性で声質を変換出来る事を示した．

謝　　辞

本研究の一部は，JSPS 科研費 22680016 と JST OngaCREST プロジェクトによる支援を受け実施したものである．

文　　献

[1] M. Morise, M. Onishi, H. Kawahara, and H. Katayose, “v. morish’ 09: A morphing-based singing design interface for vocal melodies,” Proc. ICEC, pp.185–190, 2009.

[2] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Trans. SAP, vol.6, no.2, pp.131–142, Mar. 1998.

[3] T. Toda, A.W. Black, and K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” IEEE Trans. ASLP, vol.15, no.8, pp.2222–2235, Nov. 2007.

[4] F. Villavicencio, and J. Bonada, “Applying voice conversion to concatenative singing-voice synthesis,” Proc. INTERSPEECH, pp.2162–2165, Sept. 2010.

[5] 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,” 信学技法，SP 110–297，pp.71–76，Nov. 2010.

[6] T. Toda, Y. Ohtani, and K. Shikano, “One-to-many and many-to-one voice conversion based on eigenvoices,” Proc. ICASSP, pp.1249–1252, Apr. 2007.

[7] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Many-to-many eigenvoice conversion with reference voice,” Proc. INTERSPEECH, pp.1623–1626, Sept. 2009.

[8] H. Doi, T. Toda, T. Nakano, M. Goto, and S. Nakamura, “Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system,” Proc. APSIPA ASC, Nov. 2012.

[9] H. Zen, K. Tokuda, and A.W. Black, “Statistical parametric speech synthesis,” Speech Communication, vol.51, no.11, pp.1039–1064, Nov. 2009.

[10] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, “A style control technique for HMM-based expressive speech synthesis,” IEICE Trans. Information and Systems, vol.E90-D, no.9, pp.1406–1413, Sep. 2007.

[11] M. Tachibana, T. Nose, J. Yamagishi, and T. Kobayashi, “A technique for controlling voice quality of synthetic speech using multiple regression HSMM,” Proc. INTERSPEECH, pp.2438–2441, Sept. 2006.

[12] 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，Mar. 2009.

[13] T. Nose, M. Kanemoto, T. Koriyama, and T. Kobayashi, “A style control technique for singing voice synthesis based on multiple-regression HSMM,” Proc. INTERSPEECH, pp.378–382, Aug. 2013.

[14] 木戸博，粕谷英樹，“通常発話の声質に関連した日常表現語 : 聴取評価による抽出,” 日本音響学会誌，vol.57，no.5，pp.337–344，May 2001.

[15] K. Kobayashi, H. Doi, T. Toda, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura, “An investigation of acoustic features for singing voice conversion based on perceptual age,” Proc. INTERSPEECH, pp.1057–1061, Aug. 2013.

[16] 後藤真孝，西村拓一，“AIST ハミングデータベース: 歌声研究用音楽データベース,” 情報処理学会音楽情報科学研究会研究報告，vol.2005-MUS-61-2，no.82，pp.7–12，Aug. 2005.

[17] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol.27, no.3-4, pp.187–207, Apr. 1999.

[18] K. Ohta, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Regression approaches to voice quality control based on one-to-many eigenvoice conversion,” 6th ISCA Speech Synthesis Workshop (SSW6), pp.101–106, Aug. 2007.

[19] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” Proc. INTERSPEECH, pp.1076–1079, Sept. 2008.

[20] T. Toda, T. Muramatsu, and H. Banno, “Implementation of computationally efficient real-time voice conversion,” Proc. INTERSPEECH, Sept. 2012.