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

信学技報
TECHNICAL REPORT OF IEICE.

統計的歌声声質変換における知覚年齢に基づく声質制御

小林　和弘† 　戸田　智基† 　ニュービッグ　グラム† 　サクティ　サクリアニ† 　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
E-mail: †{kazuhiro-k,tomoki,neubig,ssakti,s-nakamura}@is.naist.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．本報告では，統計的手法に基づく歌声声質変換における知覚年齢に沿った声質制御法を述べる．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となる．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音声の声質を自由に制御する事は困難である．本報告では，統計的手法に基づく歌声声質変換において，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手法を提案する．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

Voice Quality Control Based on Perceptual Age in Statistical Singing Voice Conversion

Kazuhiro KOBAYASHI†, Tomoki TODA†, Graham NEUBIG†, Sakriani SAKTI†, and Satochi NAKAMURA†

† Information Science, Nara Institute of Science and Technology
Takayama 8916–5, Ikoma, Nara, 630–0192 Japan
E-mail: †{kazuhiro-k,tomoki,neubig,ssakti,s-nakamura}@is.naist.jp

Abstract　The perceptual 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. In this paper, we describe a novel voice timbre control technique based on the perceptual age for singing voice conversion (SVC). Singers can sing expressively by controlling prosody and voice timbre, but the varieties of voices that singers can produce are limited by physical constraints. Previous work has attempted to overcome the limitation through the use of statistical 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. In this paper, we develop a technique for controlling the voice timbre based on perceptual age that maintains the singer's individuality. The experimental results show that the proposed voice timbre control method makes it possible to change the singer's perceptual age while not having an adverse effect on the perceived individuality.
Key words　singing voice conversion, perceptual age, segmental features, individuality, voice quality control.

1. はじめに

歌声は音楽において最も表現豊かな楽器の一つである．メロディーやリズムに加え言語情報である歌詞を歌声に加えることで，他の楽器にはない多彩な音楽表現を可能とする．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．

近年，歌手に代わり歌声を生成する手法として，メロディーの楽譜情報や歌詞情報から歌声を合成する Vocaloid [1] や Sinsy [2] などの歌声合成システムが盛んに利用されている．更に，人の歌声データより歌声合成システムの手動操作では表現が困難な微細な変動を推定し，歌声合成エンジンで表現する Vocalistener [3] が提案されている．歌声合成システムにより歌手の代わりに歌声を合成する代替歌唱手段は確立されたが，歌声合成システムでは歌手の声質を自由に制御することは困難である．

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 [4] がある．この手法は，異なる声質を持つ同一曲の歌声間において，声質や言語情報を表すスペクトル特徴量と音高を表す基本周波数（F_0）をそれぞれ独立に補完することで，新たな声質を持つ歌声の合成が可能である．しかし，異なる歌手間の同一曲の歌声のみが補完可能であるため利用範囲は限定される．

より柔軟な声質制御法として，統計的手法に基づく歌声声質変換（SVC: Singing Voice Conversion）[5], [6] が提案されている．この手法は，入力話者の如何なる発話内容に対しても，言語情報を保持したまま目標話者の声質へと変換する統計的手法に基づく声質変換 [7], [8] を歌声に適用したものである．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，入力歌手と目標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，学習データに存在しない楽曲に対しても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ことが可能となる．一方で，変換後の声質は目標歌手のものに限定されており，歌手の意に沿った自在な声質制御を実現するためには，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音声の声質を制御できる機能の実現が望まれる．

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 [9] 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，本研究ではその中の一つとして，歌声の知覚年齢 [10] に着目する．ここで歌声の知覚年齢とは，歌声を聞いた時に知覚されるその歌手の年齢である．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．そこで我々は，まず初めに歌声における知覚年齢と音響特徴量の関係を調査した．その結果，スペクトル特徴量や非周期成分を表す分節的特徴および F_0 やパワーなどの歌いまわしを表す韻律的特徴の両特徴量が知覚年齢に寄与しており，分節的特徴に比べ韻律的特徴がより知覚年齢に寄与することを明らかにした [10]．韻律的特徴は，楽曲に大きく依存しているものの，歌手自身の歌いまわしを変化させることで一定の制御が可能である．一方で，分節的特徴は歌手の身体的特徴に大きく依存しており，歌手が自らの分節的特徴を変化させ知覚年齢を制御する事は困難である．

本稿では，身体的制約により制御することが困難な分節的特徴に着目し，SVC において分節的特徴の変換により知覚年齢に基づく声質制御法の実現を目指す．はじめに，知覚年齢に基いて声質制御可能な重回帰 GMM（MR-GMM: Multiple-regression GMM）に基づく声質変換 [11] を SVC に適用する．さらに，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，歌手の個人性を保持しつつ知覚年齢に基づいた声質制御が可能な手法の提案を行う．なお，提案法ではメルケプストラムなどのスペクトル特徴量を用いて歌手の声質を変換するため，リアルタイム声質変換システム [12], [13] 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．

2. 関連研究

SVC において，学習データに含まれない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ために，固有声変換技術 [14], [15] を歌声に適用した手法が提案されている [16]．固有声変換の一つである多対多固有声に基づく SVC [16] 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できる．また，少量の歌声データが用意出来ない場合は，EV-GMM の適応パラメータを手動操作することで EV-GMM のパラメータを操作可能である．しかし適応パラメータ空間は人の主観軸に沿っていないため，パラメータ操作後の変換音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．本手法により，任意の目標歌手の声質による歌唱表現が可能となるが，更に豊かな音楽表現を実現するためには，特定の目標歌手の声質へと変換するだけではなく，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制御可能な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 [17] 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手法が実現している [18]．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も可能となった [19]．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 [20]．また歌声合成において，重回帰分析により，学習に用いた単一歌手内での “若さ–老い” の歌唱様式の操作法が提案されている [21]．

3. 統計的手法に基づく歌声声質変換

3.1 GMM に基づく SVC

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，$2D$ 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．

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

ここで $\mathcal{N}\left(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ 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 [8] 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，$\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ 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音声の音質劣化を緩和するため，系列内変動（GV: Global Variance）[8] を考慮する．

3.2 多対多 EV-GMM に基づく SVC

EV-GMM に基づく多対多 SVC では，任意の入力歌手の歌声を任意の目標歌手の歌声へと変換する技術である．本手法は，GMM に基づく SVC と同様に学習処理と変換処理から構成される．

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を用いて，EV-GMM により結合確率密度関数をモデル化する．参照歌手と s 番目の事前収録目標歌手のフレーム t における，$2D$ 次元の静的動的特徴量ベクトルを $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$，$\boldsymbol{Y}_t^{(s)} = [\boldsymbol{y}_t^{(s)\top}, \Delta\boldsymbol{y}_t^{(s)\top}]^\top$ とすると，EV-GMM による結合確率密度関数は以下の式で表される．

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

ここで，m 番目の分布における s 番目の事前収録目標歌手に対する平均ベクトル $\boldsymbol{\mu}_m^{(Y)}(s)$ は，次式で与えられる．

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

ここで，$\boldsymbol{e}^{(s)} = [e^{(s)}(1), \cdots, e^{(s)}(J)]^\top$ は s 番目の歌手に依存する J 次元の重みパラメータである．EV-GMM のパラメータセット $\boldsymbol{\lambda}^{(EV)}$ は，GMM のパラメータセットに加え，分布に依存した基底ベクトル $\boldsymbol{A}_m = [\boldsymbol{a}_{m,1}, \cdots, \boldsymbol{a}_{m,J}]$ およびバイアスベクトル $\boldsymbol{b}_m = [\boldsymbol{b}_{m,1}, \cdots, \boldsymbol{b}_{m,J}]$ で表される．

変換処理において参照歌手の静的動的特徴量ベクトル $\boldsymbol{X}_t$ を周辺化することで，多対多 EV-GMM の結合確率密度関数は以下の式で表される．

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

ここで，$\boldsymbol{e}^{(i)}$ および $\boldsymbol{e}^{(o)}$ は任意の入力歌手と目標歌手の重みパラメータを表す．この多対多 EV-GMM に基いて 3.1 節と同様の変換処理を行う事で，各重みパラメータに基づいた任意の入力歌手から任意の目標歌手への声質変換が可能となる．

4. SVC における知覚年齢に基づいた声質制御

本稿では，SVC において入力歌手の個人性を保持したまま，知覚年齢を操作する手法を提案する．まず MR-GMM に基づく声質変換 [11] を SVC に適用する．さらに，MR-GMM に基づく SVC に対して，多対多 EV-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操作を実現する．

4.1 MR-GMM に基づく SVC

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 同様，学習処理と変換処理により構成される．

学習処理において，一人の参照歌手と複数の事前収録目標歌手の音響特徴量の結合確率密度関数により学習する．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．

表 1　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	歌声, 知覚年齢スコア w	知覚年齢スコア w 歳の平均的な歌声	$\boldsymbol{b}_m^{(Y)}w+\overline{\boldsymbol{\mu}}_m^{(Y)}$
Modified MR-GMM	歌声, 差分知覚年齢スコア Δw	入力歌手の知覚年齢 $+\Delta w$ 歳の歌声	$\hat{\boldsymbol{\mu}}_m^{(Y)}+\boldsymbol{b}_m^{(Y)}\Delta w$

$$P\left(\boldsymbol{X}_t,\boldsymbol{Y}_t|\boldsymbol{\lambda}^{(MR)},w^{(s)}\right)$$
$$=\sum_{m=1}^{M}\alpha_m\mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t\\ \boldsymbol{Y}_t\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}_m^{(X)}\\ \boldsymbol{\mu}_m^{(Y)}(s)\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)}&\boldsymbol{\Sigma}_m^{(XY)}\\ \boldsymbol{\Sigma}_m^{(YX)}&\boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right). \quad (6)$$

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．

$$\boldsymbol{\mu}_m^{(Y)}(s)=\boldsymbol{b}_m^{(Y)}w^{(s)}+\overline{\boldsymbol{\mu}}_m^{(Y)}, \quad (7)$$

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，GMM に基づく SVC と同様に系列内変動を考慮した最尤推定 [8] により変換される．

4.2　多対多 MR-GMM に基づく SVC

参照歌手のみではなく任意の入力歌手の変換に対応するため，多対多 EV-GMM に基づく SVC [16] 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．

$$P\left(\boldsymbol{Y}_t^{(i)},\boldsymbol{Y}_t^{(o)}|\boldsymbol{\lambda}^{(MR)},w^{(i)},w^{(o)}\right)$$
$$=\sum_{m=1}^{M}P\left(m|\boldsymbol{\lambda}^{(MR)}\right)\int P\left(\boldsymbol{Y}_t^{(i)}|\boldsymbol{X}_t,m,\boldsymbol{\lambda}^{(MR)},w^{(i)}\right)$$
$$P\left(\boldsymbol{Y}_t^{(o)}|\boldsymbol{X}_t,m,\boldsymbol{\lambda}^{(MR)},w^{(o)}\right)P\left(\boldsymbol{X}_t|m,\boldsymbol{\lambda}^{(MR)}\right)\mathrm{d}\boldsymbol{X}_t$$
$$=\sum_{m=1}^{M}\alpha_m\mathcal{N}\left(\begin{bmatrix}\boldsymbol{Y}_t^{(i)}\\ \boldsymbol{Y}_t^{(o)}\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}_m^{(Y)}(i)\\ \boldsymbol{\mu}_m^{(Y)}(o)\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_m^{(YY)}&\boldsymbol{\Sigma}_m^{(YXY)}\\ \boldsymbol{\Sigma}_m^{(YXY)}&\boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right) \quad (8)$$

$$\boldsymbol{\Sigma}_m^{(YXY)}=\boldsymbol{\Sigma}_m^{(YX)}\boldsymbol{\Sigma}_m^{(XX)^{-1}}\boldsymbol{\Sigma}_m^{(XY)}, \quad (9)$$

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（7）で決定される．

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（8）の入力平均ベクトル $\boldsymbol{\mu}_m^{(Y)}(i)$ を式（7）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バイアスベクトルによ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，本稿では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．

$$\mu_m^{(Y)}(i)=\hat{\mu}_m^{(Y)}, \quad (10)$$

ここで $\hat{\mu}_m^{(Y)}$ 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．

4.3　個人性を保持した知覚年齢に基づく声質制御

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（7）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．

$$\begin{aligned}\boldsymbol{\mu}_m^{(Y)}(s)&=\boldsymbol{b}_m^{(Y)}w^{(s)}+\overline{\boldsymbol{\mu}}_m^{(Y)}\\&=\boldsymbol{b}_m^{(Y)}(w^{(i)}+\Delta w)+\overline{\boldsymbol{\mu}}_m^{(Y)}\\&=\boldsymbol{b}_m^{(Y)}w^{(i)}+\overline{\boldsymbol{\mu}}_m^{(Y)}+\boldsymbol{b}_m^{(Y)}\Delta w\\&\simeq\hat{\boldsymbol{\mu}}_m^{(Y)}+\boldsymbol{b}_m^{(Y)}\Delta w, \quad (11)\end{aligned}$$

ここで出力側の知覚年齢スコア Δw は，入力歌手の歌声の知覚年齢を基に，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，入力歌手の平均ベクトルに加え，重回帰分析より得られた代表ベクトルと差分知覚年齢スコアで決定される平均ベクトルにより表される．

5.　知覚年齢に基づく声質制御の評価

本評価では，4.2 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 1 に，MR-GMM と Modified MR-GMM の入出力の関係を示す．

5.1　実 験 条 件

本報告では，評価データベースとして，AIST ハミングデータベース：ポピュラー音楽（RWC-MDB-P-2001）[22] を用いる．本データベースは，20, 30, 40, 50 歳代の日本人男女の 75 名で構成される日本語歌詞の歌唱音声データベースである．各歌手が歌う楽曲数は 25 曲であり，各曲の長さは 20 秒程度である．

MR-GMM の学習のために，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用意する．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 の混合数はスペクトル特徴量は 128 混合，非周期成分は 32 混合である．

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（11）における差分知覚年齢スコアの設定

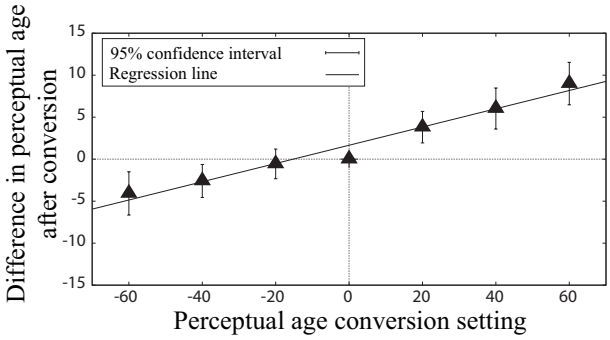

図 1　Modified MR-GMM における指定した差分知覚年齢スコアと変換音声の知覚年齢

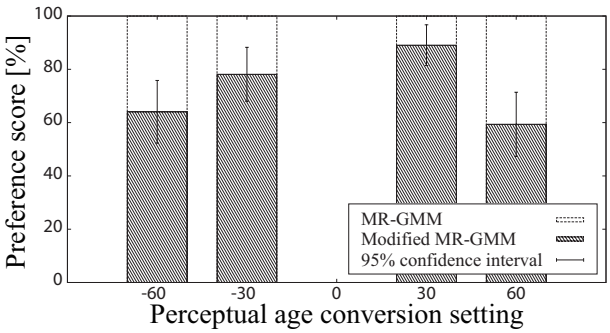

図 2　個人性に関する対比較実験結果.

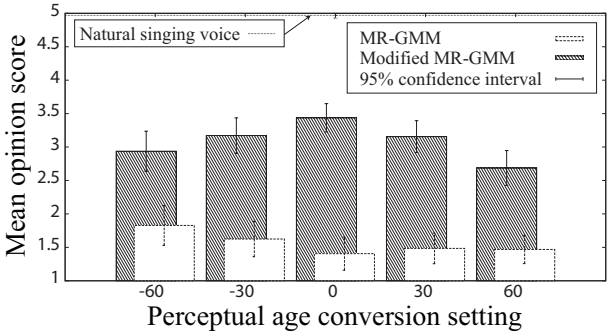

図 3　変換歌声の自然性に関する MOS 評価

は，-60, -40, -20, 0, 20, 40, 60 と変化させて変換音声を作成する．ここで差分知覚年齢スコアが 0 の時の変換音声は，声質操作を行わない同一歌手 SVC 歌声である．被験者は，ランダムな順番で提示される変換音声に対してそれぞれ知覚年齢の評価を行う．

Modified MR-GMM と MR-GMM を XAB テストにより個人性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（11）における差分知覚年齢スコアの設定は，-60, -30, 30, 60 と変化させて変換音声を作成する．式（7）における MR-GMM の知覚年齢スコアの設定は，前実験で得られた同一歌手 SVC 歌声に対する知覚年齢評価結果の歌手当たりの平均値より，± 30, 60 と変化させた値を用いる．各評価歌手の同一歌手 SVC 歌声を先に提示し，MR-GMM と Modified MR-GMM をランダムな順で再生し，どちらが同一歌手 SVC 歌声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 に対して変換音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，自然音声および Modified MR-GMM，MR-GMM の変換音声をランダムな順番で提示する．被験者は，それぞれの変換音声に対して “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．

サンプリング周波数は，16khz である．音響特徴量として，STRAIGHT 分析 [23] により抽出された，メルケプストラム係数の 1 次元から 24 次元を用いる．音源情報は，F_0 と 0–1, 1–2, 2–4, 4–6, 6–8 kHz の 5 周波数帯に平均された非周期成分を用いる．フレームシフト長は 5ms である．

5.2 実験結果

図 1 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．この線形性の関係は, [10] で観測された，分節的特徴の影響とほぼ同様の傾向で変化している事が確認出来る．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モデル化出来ていることがわかる．

図 2 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．差分知覚年齢スコアの設定を大きく変化させることで，Modified MR-GMM の個人性は，徐々に失われていく事がわかる．しかし，MR-GMM に比べ全ての差分知覚年齢スコアの設定において，歌手の個人性は多く含まれており Modified MR-GMM の有効性が確認される．

図 3 に，変換音声の自然性に関する 5 段階評価 MOS の評価結果を示す．図 2 と同様に，差分知覚年齢スコアの設定値を大きくすると変換音声の自然性が下がる傾向がある事がわかる．しかし，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．

図 1，2，3 より，Modified MR-GMM は統計的手法に基づく SVC において，歌手の個人性を保ったまま知覚年齢に基づいた声質制御を可能し，MR-GMM に比べ高い自然性で声質を変換出来ることがわかる．

6. 結　　論

本報告では，統計的手法に基づく歌声声質変換において知覚年齢に基づいた声質制御を実現するために，歌手の個人性を保持した知覚年齢変換を提案した．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，変換音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] H. Kenmochi, and H. Ohshita, “VOCALOID – Commericial singing synthesizer based on sample concatenation,” Proc. INTERSPEECH, pp.4011–4012, Aug. 2007.

[2] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, “Recent development of the HMM-based singing voice synthesis system - Sinsy,” SSW7, pp.211–216, Sept. 2010.

[3] T. Nakano, and M. Goto, “Vocalistener2: A singing synthesis sytem able to mimic a user’s singing in terms of voice timbre changes as well as pitch and dynamics,” Proc. ICASSP, pp.453–456, May 2011.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

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

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and system straight,” Proc. MAVEBA, Sept. 2001.