知覚年齢に沿った歌声声質制御のための音響特徴量の調査

小林 和弘[1,a)]　土井 啓成[1,b)]　戸田 智基[1,c)]　中野 倫靖[2,d)]　後藤 真孝[2,e)]
ニュービッグ グラム[1,f)]　サクリアニ サクテイ[1,g)]　中村 哲[1,h)]

概要：歌声は，歌詞，メロディー，声質などを駆使することで，多様な表現を生み出すことが可能である．しかし，歌手は自身の身体的制約を超えた歌声を発することは困難である．近年，この身体的制約を超えた歌唱を実現する技術として，統計的手法に基づく歌声声質変換が提案されている．この手法は，個々の歌手の声質を別の歌手の声質へと自由に変換することができるため，新たな音楽表現を可能とし，音楽制作を活性化させると期待される．より操作性に優れた歌声声質変換として，直感的に理解しやすい声質制御技術を実現できれば，さらに豊かな音楽表現が可能となる．本研究では，直感的な理解が容易であり，声質操作の対象となり得る要因の一つとして，歌声の知覚年齢に着目する．本稿では，知覚年齢の制御を可能とする声質制御技術の確立を目指し，歌声の知覚年齢に寄与する音響特徴量の調査を行う．音声分析合成処理や声質変換処理により，各音響特徴量が知覚年齢に与える影響を個別に評価する．実験結果より，分節的特徴に比べ，韻律的特徴が知覚年齢により大きく寄与することを示す．

1. はじめに

歌声は，言語情報である歌詞に対して，メロディーやリズムを与えることで，多様な表現を生み出すことができる．さらには，歌手の技量に依るものの，声質に関しても，声帯や調音器官を巧みに操ることで，変化させることが可能である．しかしながら，声質は身体的な制約が大きく反映されるため，個々の歌手が表現できる声質は限定される．身体的制約を超え，歌手の意に沿った自由な声質制御が可能となれば，更に豊かな音楽表現を生み出すことができると期待される．

歌声において，声質を変化させる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．この手法は，異なる声質を持つ同一曲の歌声間において，スペクトル包絡や基本周波数（F_0）などの音響特徴量を各々独立に補間することで，新たな声質を持つ歌声を生成する．一方で，補間対象として同一曲を必要とするため，声質を変換した歌声を生成できるのは，その曲に限定される．

より柔軟に歌声の声質を変化させる手法として，ある話者から異なる話者へと声質を変換する統計的手法に基づく声質変換技術 [2], [3] の歌声への適用が研究されている [4], [5]．この手法は，変換元である源歌手と変換先である目標歌手による同一曲の歌声（パラレルデータ）を学習データとして使用し，個々の音響特徴量に対する変換モデルを事前に学習する．代表的な変換モデルとして，源歌手と目標歌手の音響特徴量の結合確率密度関数をモデル化した混合正規分布モデル（GMM: Gaussian Mixture Model）が用いられる．学習された GMM を用いることで，源歌手による如何なる曲の歌声に対しても，目標歌手の歌声へと声質を変換することが可能となる．さらに，学習データに含まれない源歌手および目標歌手の間での歌声声質変換を実現するために，固有声変換技術 [6], [7] を歌声へと適用した手法も提案されている [8]．この手法では，多数の歌手と一人の参照歌手との間のパラレルデータセットを用いて，固有声混合正規分布モデル（EV-GMM: Eigenvoice GMM）の学習を行う．任意の源歌手および目標歌手に対する変換モデルは，各歌手による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．本手法により，個々の歌手は，任意の目標歌手の声質による歌唱が可能となるが，さらに豊かな音楽表現を可能とするためには，目

[1] 奈良先端科学技術大学院大学
Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)
[2] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)
[a)] kazuhiro-k[at]is.naist.jp
[b)] hironori-d[at]is.naist.jp
[c)] tomoki[at]s.naist.jp
[d)] t.nakano[at]aist.go.jp
[e)] m.goto[at]aist.go.jp
[f)] neubig[at]is.naist.jp
[g)] ssakti[at]is.naist.jp
[h)] s-nakamura[at]is.naist.jp

標歌手の声質へと変換するのではなく，個々の歌手が自身の思い描く所望の声質へと変換する声質制御技術の構築が望まれる．

統計的パラメトリック音声合成の研究において，声質の手動設定を可能とする技術が提案されている．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技術 [9] においては，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像を内包した重回帰 HMM を用いることで，合成音声の発話様式を手動制御する機能を実現している [10]．さらに，”暖かい”や”冷たい”などの声質表現語対 [11] 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，合成音声の声質を手動で制御することも可能となる [12]．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 [13]．テキスト音声合成のみでなく，声質変換においても，声質表言語対に対する主観評価値に基づく声質制御法が提案されている [14]．主に話声に対する研究が盛んに行われているが，これらの技術を歌声声質変換に対しても適用することで，歌声においても直感的な声質制御が実現できると期待される．

歌声の声質制御を実現する上で，話声における声質表言語対のように，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，歌声を聞いた時に感じるその歌手の年齢である．知覚年齢に沿った声質制御が実現すれば，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．話声では，スペクトル包絡パラメータとパワー情報，モーラ数などの韻律的特徴を用いて知覚年齢に基づく若年層と高齢層の話者分類を行う手法が提案されている [15]．また，話者の年齢が高くなるにつれて音源の雑音成分が増すなど，実際の年齢の遷移に伴う音響特徴量の変化についても調査されている [11]．一方で，歌声に対しては，このような研究はあまり行われておらず，知覚年齢と実年齢の対応や年齢変化に伴う音響特徴量の変化，知覚年齢に大きく影響を与える音響特徴量などは、依然としては明らかになっていない．

本報告では，知覚年齢に基づく声質制御法を実現するための第１段階として，知覚年齢に寄与する音響特徴量の調査を行う．多数歌手による歌声データを用いて，１）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の調査，および，２）歌声声質変換における知覚年齢に寄与する音響特徴量の調査を行う．実験結果から，分節的特徴に比べ，韻律的特徴が知覚年齢により大きく寄与することを示す．

2. 統計的手法に基づく歌声声質変換

統計的手法に基づく歌声声質変換（SVC: Singing Voice Conversion）は，歌手の歌声を異なる歌手の歌声へと変換する技術である．SVC は学習処理と変換処理で構成される．図１，２にそれぞれ学習処理と変換処理を示す．

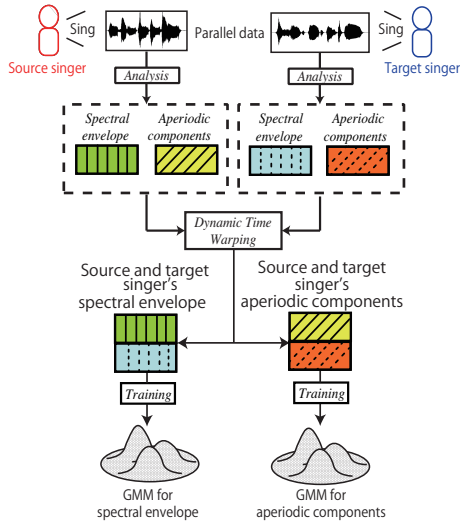

図 1 統計的手法に基づく歌声声質変換の学習処理

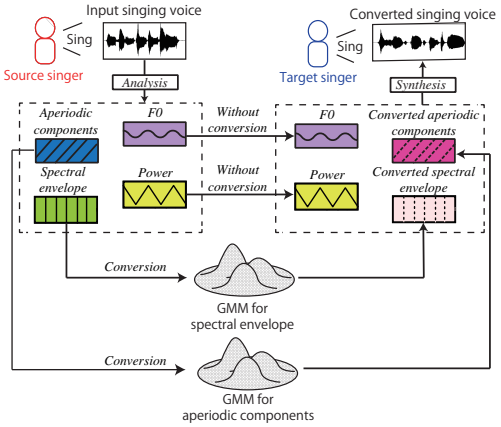

図 2 統計的手法に基づく歌声声質変換の変換処理

学習処理では，話声の声質変換と同様に，源歌手と目標歌手のパラレルデータセットより音響特徴量を抽出し，GMM により結合確率密度関数をモデル化する．源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 における源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，同フレームの源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．パラレルデータセットに対して，動的時間伸縮によ

表 1 各合成歌声に内包する音響特徴量

合成手法	分析再合成（w/ AC）	非周期成分無し分析再合成（w/o AC）	同一歌手 SVC	SVC
メルケプストラム	源歌手	源歌手	源歌手	目標歌手
非周期成分	源歌手	未使用	源歌手	目標歌手
パワー，F_0，継続長	源歌手	源歌手	源歌手	源歌手

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，源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．源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．

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

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）[3] を考慮する．

3. 知覚年齢に寄与する音響特徴量の調査

SVC[5], [8] では, GMM を用いた変換処理を施す音響特徴量として，メルケプストラムや非周期成分（AC: Aperiodic Components）[16] などの分節的特徴を主な対象とする．これらの音響特徴量が歌声の知覚年齢に大きく影響を与えるのであれば，声質表言語対に対する主観評価値に基づく声質制御技術 [14] を SVC に導入することで，歌声の知覚年齢操作が実現できると予想される．さらには，リアルタイム声質変換技術 [17], [18] も組み合わせることで，歌声の知覚年齢のリアルタイム操作を用いた新たな歌唱表現を実現できる可能性がある．

一方で，歌声の知覚年齢が，分節的特徴ではなく，パワーパターンや F_0 パターン，継続長などの韻律的特徴の影響を大きく受けるのであれば，これらの特徴量を制御する必要がある．韻律的特徴を高精度に変換するためには，HMM 音声合成に基づく声質制御技術 [10], [12] のように，コンテキスト情報を利用して音響特徴量をモデル化する枠組みが有効である．この場合，オフライン処理による歌声の知覚年齢制御の実現が見込まれる．一方で，SVC で実現が期待されるリアルタイム知覚年齢操作を用いた歌唱表現において，高精度な韻律的特徴の変換を行うのは本質的に困難となる．そのため，SVC による分節的特徴の変換に加え，歌手自身が韻律的特徴を制御した歌唱を行う必要がある．

上記のように，変換処理を施す音響特徴量に応じて，実現が見込まれる技術は変化するため，歌声の知覚年齢を操作する上でどの音響特徴量を変換する必要があるかを調査する．知覚年齢に寄与する音響特徴量を調査するために，自然歌声の知覚年齢と 3.1 節から 3.4 節に示す合成歌声の知覚年齢の比較を行う．表 1 に，各合成手法と合成歌声の特徴を示す．

3.1 分析再合成ひずみによる影響

分析再合成は，歌声声質変換や HMM に基づく歌声合成において欠かせない処理である．そこで，分析再合成により生じるひずみが歌声の知覚年齢に与える影響を調査する．自然歌声から，音響特徴量としてメルケプストラム，F_0，非周期成分を抽出し，音響特徴量の変形処理は一切施さずに波形合成を行う．本報告では，上記処理により得られる合成歌声を，分析再合成歌声（w/ AC）とする．高精度な分析合成法として，STRAIGHT[19] を用い，波形合成時における音源モデルには非周期成分に基づく混合励振源 [20] を用いる．

3.2 非周期成分の影響

音源の雑音成分は，話声において話者の年代により変化する傾向が観測されている [11]．そこで，音源の雑音成分を捉える音響特徴量として，非周期成分が歌声の知覚年齢に与える影響を調査する．STRAIGHT を用いて，自然歌声からメルケプストラムと F_0 を抽出する．合成時には，混合励振源ではなく，簡易な位相制御を施したパルス列で構成される有声音源 [19] と雑音源を切り替えることで音源信号を生成する．得られた合成歌声を非周期成分無し分析再合成歌声（w/o AC）とする．3.1 節で述べた分析再合成歌声（w/ AC）と，分析再合成歌声（w/o AC）の知覚年齢スコアを比較することで，非周期成分が知覚年齢に与える影響を調査する．

3.3 統計的手法に基づく声質変換による影響

SVC や HMM に基づく歌声合成においては，統計処理による変換誤差の影響は避けられない．本報告では，SVC を対象とし，GMM に基づく変換処理により生じる変換誤差の影響について調査する．SVC では，変換処理を通して，例えばスペクトル包絡の詳細な構造などは除去される傾向がある．このような変換処理により失われる音響特徴量が，歌声の知覚年齢に与える影響を調査するために，ある歌手から同じ歌手への SVC（同一歌手 SVC）を行う．同一歌手 SVC を実現するためには，結合確率密度関数 $P(\boldsymbol{X}_t, \boldsymbol{X}'_t|\boldsymbol{\lambda})$ を得る必要がある．ここで $\boldsymbol{X}_t$ と $\boldsymbol{X}'_t$ は同一歌手の音響特徴量ベクトルを表し，お互いに異なるものの，どちらも同一の確率密度関数に従う（すなわち，$P(\boldsymbol{X}_t|\boldsymbol{\lambda}) = P(\boldsymbol{X}'_t|\boldsymbol{\lambda})$）．このような結合確率密度関数をモデル化する GMM を学習するためには，例えば，同一歌手が同じ曲を複数回歌唱す

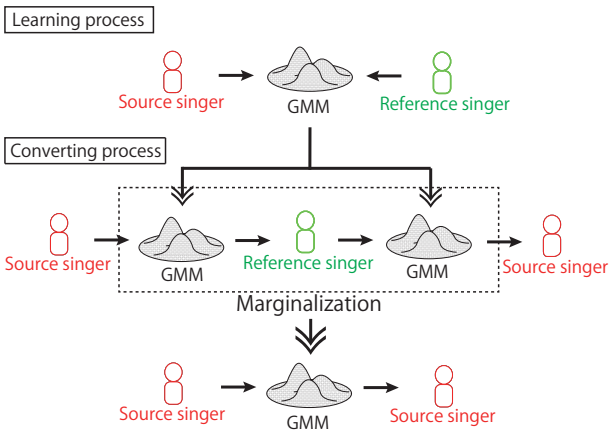

図 3 同一歌手 SVC の枠組み

ることで得られる歌声データを用いるという方法も考えられるが，本報告では，より容易な方法として，多対多固有声変換 [6], [7], [8] で用いられている枠組みを応用する．

図 3 に同一歌手 SVC の枠組みを示す．2 節の SVC と同様に，源歌手と異なる歌手である参照歌手のパラレルデータを用いて，GMM を学習する．この GMM を用いることで，源歌手の音響特徴量から参照歌手の音響特徴量への変換処理と，それとは逆に参照歌手から源歌手への変換処理を実現できる．これらの変換処理を繋ぎ合わせ，かつ中間結果である参照歌手の音響特徴量を周辺化することで，同一歌手 SVC を実現する．ここで，源歌手と参照歌手に対する GMM でモデル化される結合確率密度関数を $P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda})$ とし，$\boldsymbol{X}_t$ と $\boldsymbol{Y}_t$ を各々源歌手の音響特徴量ベクトルと参照歌手の音響特徴量ベクトルとする．この時，同一歌手 SVC で用いられる結合確率密度関数は次式の GMM により与えられる．

$$P\left(\boldsymbol{X}_t, \boldsymbol{X}'_t|\boldsymbol{\lambda}\right) = \sum_{m=1}^{M} P\left(m|\boldsymbol{\lambda}\right) \int P\left(\boldsymbol{X}_t|\boldsymbol{Y}_t, m, \boldsymbol{\lambda}\right) P\left(\boldsymbol{X}'_t|\boldsymbol{Y}_t, m, \boldsymbol{\lambda}\right) P\left(\boldsymbol{Y}_t|m, \boldsymbol{\lambda}\right) \mathrm{d}\boldsymbol{Y}_t$$
$$= \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{X}'_t\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(X)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XYX)} \\ \boldsymbol{\Sigma}_m^{(XYX)} & \boldsymbol{\Sigma}_m^{(XX)}\end{bmatrix}\right), \quad (3)$$

$$\boldsymbol{\Sigma}_m^{(XYX)} = \boldsymbol{\Sigma}_m^{(XY)} \boldsymbol{\Sigma}_m^{(YY)^{-1}} \boldsymbol{\Sigma}_m^{(YX)}, \quad (4)$$

この GMM により，2 節と同様の変換処理で同一歌手 SVC による変換歌声を得ることができる．得られた変換歌声と分析再合成歌声（w/ AC）を比較することで，SVC における変換誤差が知覚年齢に与える影響を調査する．

3.4 韻律的特徴と分節的特徴の影響

音響特徴量の内，分節的特徴と韻律的特徴のどちらが知覚年齢に大きく寄与しているかを調査する．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，源歌手から目標歌手への変換歌声を合成する．結果，得られる変換歌声は，源歌手の持つ 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持つメルケプストラム，AC 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，目標歌手の同一歌手 SVC による変換歌声の知覚年齢を比較することで，どちらの音響特徴量がより知覚年齢に寄与するかを明らかにする．

4. 実験的評価

4.1 実験条件

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．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）[21] 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．20 代男性 1 名の被験者が，全楽曲に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，表 1 に示す各合成歌声と自然歌声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，各合成歌声と自然歌声に対し知覚年齢スコアを付与する．被験者への負担を減らすため，歌手の実年齢と歌声の知覚年齢スコアの相関が最も高い P039 を評価楽曲とする．さらに実年齢と知覚年齢の相関が高い男女を実年齢の各年代別に 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．

歌声声質変換及び HMM 音声合成において，知覚年齢に沿った声質制御は，歌手の話者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，SVC 歌声の持つ話者性が，韻律的特徴か分節的特徴のどちらに多く反映されているかを振り分けテストにより評価する．

表 1 において，同一話者 SVC による合成歌声と比較し，SVC による合成歌声は，源歌手から目標歌手へと分節的特徴を変換したもの，もしくは，目標歌手から源歌手へと韻律的特徴を変換したものとみなすことができる．これらの合成歌声を用いて，知覚年齢変換処理における話者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 による合成歌声（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 による合成歌声と，各セットにおける個々の評価歌手の同一歌手 SVC による合成歌声を比較し，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 16kHz である．音響特徴量として STRAIGHT で抽出されたメルケプストラム係数の 1 次元

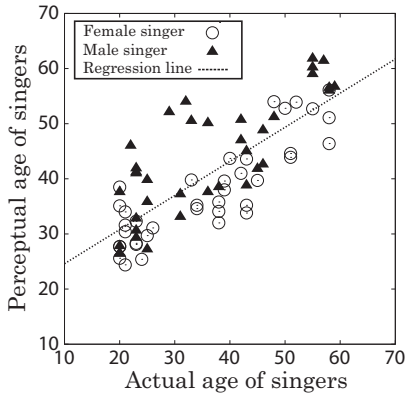

図 4 歌手の実年齢と知覚年齢スコアの相関図

表 2 自然歌声と各合成歌声の知覚年齢スコアの差

合成歌声の種類	差分の平均値	標準偏差	相関係数
分析再合成歌声（w/ AC）	0.77	3.57	0.96
分析再合成歌声（w/o AC）	0.44	3.58	0.96
同一歌手 SVC 歌声	-0.5	7.25	0.85

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．

同一歌手 SVC 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，メルケプストラム及び非周期成分を変換するための GMM は，各グループ内において評価歌手の総当りペアに対して学習及び変換を行う．混合数は，各評価歌手ペアにおいて，最適な値を用いる．

4.2 実験結果

図 4 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．全体の相関係数は 0.79 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．なお，女性の相関係数は 0.86 であり，男性の相関係数は 0.80 である．

表 2 に自然歌声と各合成歌声の知覚年齢スコアの平均値の差分と，標準偏差及び相関係数を示す．分析再合成歌声（w/ AC）の知覚年齢スコアと自然歌声の知覚年齢スコアの差分の平均値は 1 歳未満と小さい．この結果より，分析再合成ひずみが知覚年齢に与える影響は非常に小さいことがわかる．同様に，分析再合成歌声（w/o AC）の知覚年齢スコアと分析再合成歌声（w/ AC）の知覚年齢スコアの差分の平均値の差は小さい．これより，非周期成分が歌声の知覚年齢に与える影響は，非常に小さいことがわかる．一方，同一歌手 SVC 歌声の知覚年齢スコアと自然歌声の知覚年齢スコアには，わずかな差が発生する．このことから，GMM を用いた変換処理に伴う変換誤差は，知覚年齢に多少なりとも影響を与えることが分かる．しかしながら，知覚年齢スコアの差の平均値は小さく，相関係数も高いため，変換後の音響特徴量においても知覚年齢に影響を与える情報は概ね保持されていると考えられる．

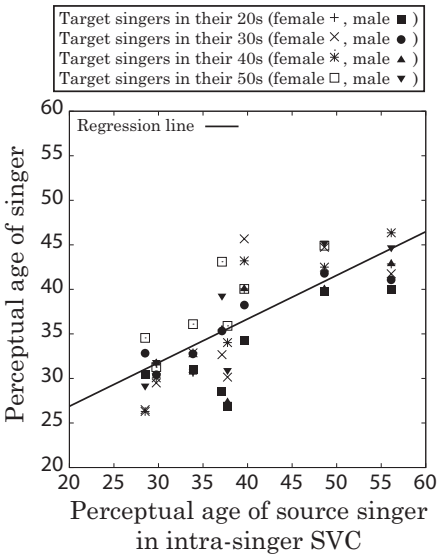

図 5 同一歌手 SVC 歌声と SVC 歌声の知覚年齢の対応図（横軸を源歌手の同一歌手 SVC 歌声の知覚年齢スコアにした場合）

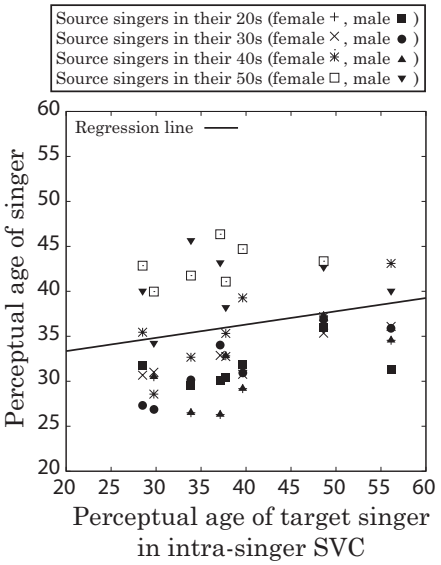

図 6 同一歌手 SVC 歌声と SVC 歌声の知覚年齢の対応図（横軸を目標歌手の同一歌手 SVC 歌声の知覚年齢スコアにした場合）

図 5，6 に，同一歌手 SVC 歌声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 5 は，横軸を源歌手の同一歌手 SVC 歌声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 6 は，横軸を目標歌手の同一歌手 SVC 歌声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 3 に，SVC において韻律的特徴もしくは分節的特徴の

表 3 SVC における話者性の評価

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

変換を行った際に生じる話者性の変化に対する評価結果を示す．表は，源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，源歌手の同一歌手 SVC 歌声（韻律的特徴が一致）に似ていると判断された場合の確率，目標歌手の同一歌手 SVC 歌声（分節的特徴が一致）に似ていると判断された場合の確率，源歌手と目標歌手以外の同一歌手 SVC 歌声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，歌手の話者性は，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５，６の結果と同様の傾向であることから，話者性と知覚年齢の相関は高いといえる．これは，変換時に目標話者への変換を行っているためであり，妥当な結果である．話者性をできる限り保存したまま知覚年齢を制御するためには，話者性と知覚年齢の影響を分離し，話者非依存の知覚年齢変換処理を実現する必要があるといえる．

5. 結論

本稿では，歌声において知覚年齢に寄与する音響特徴量の調査を行った．様々な合成歌声の知覚年齢の比較を行うことで，知覚年齢に寄与する音響特徴量の調査を行った．実験結果より，１）分析再合成や歌声声質変換における処理ひずみが知覚年齢に及ぼす影響は小さく，２）韻律的特徴は分節的特徴に比べ知覚年齢に大きく寄与することが分かった．今後は，話者性を保持した知覚年齢操作を可能とする歌声声質制御技術の研究を行う．

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．

参考文献

[1] 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. ICASSP*, pp. 3905–3908 (2009).

[2] Stylianou, Y., Cappé, O. and Moulines, E.: Continuous Probabilistic Transform for Voice Conversion, *IEEE Trans. SAP*, Vol. 6, No. 2, pp. 131–142 (1998).

[3] Toda, T., Black, A. W. and Tokuda, K.: Voice conversion based on maximum likelihood estimation of spectral parameter trajectory, *IEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222–2235 (2007).

[4] Villavicencio, F. and Bonada, J.: Applying voice conversion to concatenative singing-voice synthesis, *Proc. INTERSPEECH*, pp. 2162–2165 (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，電子情報通信学会技術研究報告，Vol. SP2010 69-87, No. 297, pp. 71–76 (2010).

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

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

[8] Doi, H., Toda, T., Nakano, T., Goto, M. and Nakamura, S.: 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* (2012).

[9] Zen, H., Tokuda, K. and Black, A. W.: Statistical parametric speech synthesis, *Speech Communication*, Vol. 51, No. 11, pp. 1039–1064 (2009).

[10] Nose, T., Yamagishi, J., Masuko, T. and Kobayashi, T.: A Style Control Technique for HMM-Based Expressive Speech Synthesis, *IEICE Trans. Information and Systems*, Vol. E90-D, No. 9, pp. 1406–1413 (2007).

[11] Kasuya, H., Yoshida, H., Ebihara, S. and Mori, H.: Longitudinal Changes of Selected Voice Source Parameters, *Proc. INTERSPEECH*, pp. 2570–2573 (2010).

[12] Tachibana, M., Nose, T., Yamagishi, J. and Kobayashi, T.: A technique for controlling voice quality of synthetic speech using multiple regression HSMM, *Proc. INTERSPEECH*, pp. 2438–2441 (2006).

[13] 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).

[14] Ohta, K., Toda, T., Ohtani, Y., Saruwatari, H. and Shikano, K.: Adaptive voice-quality control based on one-to-many eigenvoice conversion, *Proc. INTERSPEECH*, pp. 2158–2161 (2010).

[15] Minematsu, N., Sekiguchi, M. and Hirose, K.: Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers, *Proc. ICASSP*, pp. 137–140 (2002).

[16] Kawahara, H., Estill, J. and Fujimura, O.: Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and system STRAIGHT, *Proc. MAVEBA* (2001).

[17] Muramatsu, T., Ohtani, Y., Toda, T., Saruwatari, H. and Shikano, K.: Low-Delay Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory, *Proc. INTERSPEECH*, pp. 1076–1079 (2008).

[18] Toda, T., Muramatsu, T. and Banno, H.: Implementation of computationally efficient real-time voice conversion, *Proc. INTERSPEECH* (2012).

[19] Kawahara, H., Masuda-Katsuse, I. and Cheveigné, A.: 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 (1999).

[20] Ohtani, Y., Toda, T., Saruwatari, H. and Shikano, K.: Maximum Likelihood Voice Conversion Based on GMM with STRAIGHT Mixed Excitation, *Proc. INTERSPEECH*, pp. 2266–2269 (2006).

[21] 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 (2005).