統計的歌声声質変換における知覚年齢に沿った声質制御*

☆小林 和弘，戸田 智基，Graham Neubig，Sakriani Sakti，中村 哲（奈良先端大・情報）

1　はじめに

歌声は音楽を形成する上で重要な要素の 1 つであり，人は歌声の音高や音色を巧みに操作する事で，多様な歌唱表現を生み出す事が可能である．一方で，個人の持つ声質は身体的特徴により大きく制限されており，身体的特徴を超えた声色での歌唱は困難である．近年，この身体的制約を超える声質制御法として，統計的手法に基づく歌声声質変換（SVC: Singing Voice Conversion）が提案され [1]，歌手は多様な声質での歌唱が可能となった．しかし，人の主観に基づく直感的な声質制御を実現するまでには至っていない．

本稿では，主観的情報の 1 つである「知覚年齢」に着目し，知覚年齢に沿った声質制御を実現する．まず，話し声において有効性が確認されている重回帰混合正規分布モデル（multiple-regression Gaussian mixture model: MR-GMM）に基づく声質変換法 [2] を，SVC に適用する．さらに，歌手の個人性を保持した声質制御を実現するための手法を提案する．実験結果より，歌手の個人性を保持しつつ知覚年齢に基づく歌声声質制御が可能であることを示す．

2　重回帰混合正規分布モデルに基づく声質制御

MR-GMM に基づく声質制御は，入力話者の声質を，話者の身体的特徴や声質を数値化した声質表現語スコアに基いて，所望の声質へと変換する技術である [2]．一人の参照話者と複数の事前収録目標話者が同一文セットを発声したパラレルデータを用いて，次式の MR-GMM を学習する．

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

ここで，$\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ 及び $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^{(s)\top}, \Delta\boldsymbol{y}_t^{(s)\top}\right]^\top$ は，参照話者と s 番目の事前収録目標話者の静的・動的特徴量ベクトルを表す．$\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は平均ベクトル $\boldsymbol{\mu}$ 及び共分散行列 $\boldsymbol{\Sigma}$ を持つ正規分布を表す．MR-GMM の混合数は M であり，m は分布番号を示す．m 番目の分布における s 番目の事前収録目標話者に対する平均ベクトル $\boldsymbol{\mu}_m^{(Y)}(s)$ は，次式で与えられる．

$$\boldsymbol{\mu}_m^{(Y)}(s) = \boldsymbol{B}_m^{(Y)}\boldsymbol{w}^{(s)} + \overline{\boldsymbol{\mu}}_m^{(Y)} \quad (2)$$

ここで，$\boldsymbol{B}_m^{(Y)}$ 及び $\overline{\boldsymbol{\mu}}_m^{(Y)}$ は，声質表現語スコアに対する代表ベクトルセット及びバイアスベクトルを表す．また，$\boldsymbol{w}^{(s)}$ は，s 番目の事前収録目標話者の声質表現語スコアを表し，声質制御者の主観に基づいて人手で与える．

変換処理では，所望の声質表現語スコア $\boldsymbol{w}$ を用いて得られる MR-GMM に基づき，最尤系列変換法 [3] により，参照話者の音声を所望の声質を持つ音声へと変換する．

3　知覚年齢に沿った歌声声質制御

[4] において，韻律的特徴及び分節的特徴の両音響特徴量が知覚年齢に与える影響を調査し，両特徴量とも知覚年齢に影響を与えること，韻律的特徴の方が知覚年齢に大きく寄与するが個人性にも大きな影響を与えること，が報告されている．本稿では，分節的特徴は韻律的特徴と比較して歌手が制御できる範囲が狭い点に着目し，分節的特徴の変換により，歌手の身体的制約を超えた声質制御の実現に取り組む．その際に，歌手の個人性を保持した声質制御の実現を目指す．

3.1　多対多 MR-GMM に基づく SVC

知覚年齢に沿った歌声声質制御を実現するために，MR-GMM に基づく声質制御 [2] を多対多 SVC[1] に適用する．多対多 MR-GMM は以下の式で表される．

$$\begin{aligned} &P\left(\boldsymbol{Y}_t^{(i)}, \boldsymbol{Y}_t^{(o)}|\boldsymbol{\lambda}^{(MR)}, w^{(i)}, w^{(o)}\right) \\ &= \sum_{m=1}^{M} P\left(m|\boldsymbol{\lambda}^{(MR)}\right) \int P\left(\boldsymbol{Y}_t^{(i)}|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(MR)}, w^{(i)}\right) \\ &\quad P\left(\boldsymbol{Y}_t^{(o)}|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(MR)}, w^{(o)}\right) P\left(\boldsymbol{X}_t|m, \boldsymbol{\lambda}^{(MR)}\right) \mathrm{d}\boldsymbol{X}_t \\ &= \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{Y}_t^{(i)} \\ \boldsymbol{Y}_t^{(o)}\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(Y)}(i) \\ \boldsymbol{\mu}_m^{(Y)}(o)\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(YY)} & \boldsymbol{\Sigma}_m^{(YXY)} \\ \boldsymbol{\Sigma}_m^{(YXY)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right) \quad (3) \end{aligned}$$

$$\boldsymbol{\Sigma}_m^{(YXY)} = \boldsymbol{\Sigma}_m^{(YX)}{\boldsymbol{\Sigma}_m^{(XX)}}^{-1}\boldsymbol{\Sigma}_m^{(XY)} \quad (4)$$

ここで，$w^{(i)}$ 及び $w^{(o)}$ は，入力歌手の知覚年齢スコアおよび目標歌手の知覚年齢スコアを表し，入力ベクトルおよび出力ベクトルは対応する知覚年齢スコアにより，次式により表される．

$$\boldsymbol{\mu}_m^{(Y)}(s) = \boldsymbol{b}_m^{(Y)}w^{(s)} + \overline{\boldsymbol{\mu}}_m^{(Y)} \quad (5)$$

ここで，$\boldsymbol{b}_m^{(Y)}$ 及び $\overline{\boldsymbol{\mu}}_m^{(Y)}$ は知覚年齢スコアに対応する代表ベクトル及びバイアスベクトルを表す．

声質制御対象となる歌手に対して，多対多 MR-GMM を適用する．歌手制御対象歌手の知覚年齢スコアに基づき，入力平均ベクトルを式 (5) で与えることも可能であるが，モデル化の精度は下がる．一方で，声質制御対象歌手の十分な量の歌声データが入手可能であれば，式 (5) を用いずに，入力平均ベクトル自体を最大事後確率推定することも可能である．本稿では，理想的な状況として，声質制御対象歌手と MR-GMM 学習時に用いた参照歌手 1 名とのパラレルデータが入手可能である場合を想定し，入力平均ベクトルの最尤推定を行う．ここで，最尤推定された入力平均ベクトルを $\boldsymbol{\mu}_m^{(Y)}(s)$ とする．なお，$\boldsymbol{\mu}_m^{(Y)}(o) = \boldsymbol{\mu}_m^{(Y)}(s)$ とすることで，同一の入出力歌手で変換した変換音声も生成可能である．本稿では，この変換音声を同一歌手 SVC 歌声と呼ぶ．

3.2　個人性を保持する歌声声質制御

多対多 MR-GMM に基づく SVC では，出力側の知覚年齢スコアを指定することで，所望の知覚年齢を

---

*Voice Quality Control Based on Perceptual Age in Singing Voice Conversion, by KOBAYASHI, Kazuhiro, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani, NAKAMURA, Satoshi（NAIST）

持つ声質への歌声声質変換が可能となる．しかし，式 (5) により得られる出力平均ベクトルは，複数の事前収録目標歌手の平均的な声質の特徴を表現するものとなり，特定の歌手の声質を表現していない．そのため，声質制御対象歌手の個人性を保ちながら，知覚年齢を制御することはできない．

個人性を保持した知覚年齢制御を実現するために，出力平均ベクトルの表現形式を変更する．式 (5) では，バイアスベクトルは全事前収録目標歌手の平均的な声質を表現しており，代表ベクトルは知覚年齢の変化に伴う平均ベクトルの変化を表す．これに対して，次式の通り，バイアスベクトルを声質制御対象歌手の平均ベクトル $\hat{\boldsymbol{\mu}}_m^{(Y)}$ へと置き換える．

$$\boldsymbol{\mu}_m^{(Y)}(o) = \hat{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{b}_m^{(Y)} \Delta w \qquad (6)$$

ここで，Δw は声質制御対象歌手の知覚年齢を変化させる差分知覚年齢スコアである．これにより，全事前収録目標歌手の平均的な声質を中心とした部分空間ではなく，声質制御対象歌手の声質を中心とした部分空間により，出力平均ベクトルが表現される．

4 実験的評価

4.1 実験条件

歌唱データとして，AIST ハミングデータベース：ポピュラー音楽（RWC-MDB-P-2001）日本語歌詞，サビパート [5] を用いる．評価楽曲は No.39 とする．MR-GMM の学習において，参照歌手として実年齢が 20 代の女性 1 名を用い，事前収録目標歌手として実年齢が 20 代，30 代，40 代，50 代の女性 27 名，男性 27 名を用いる．評価用目標歌手として，事前収録目標歌手に含まれない 16 名（実年齢が 20 代，30 代，40 代，50 代の男女各 2 名ずつ）を評価歌手（声質制御対象歌手）として用いる．被験者は 20 代男性 8 名である．

スペクトル包絡パラメータとして，STRAIGHT 分析 [6] によって得られるスペクトル包絡から算出される 1 次から 24 次のメルケプストラム係数を使用する．シフト長は 5 ms，サンプリング周波数は 16 kHz とする．音源特徴量は，F_0 と 5 周波数帯域における平均非周期成分を使用する．スペクトル包絡と非周期成分の GMM の混合数はそれぞれ 128，32 である．

知覚年齢に基づく歌声声質制御の精度を評価するため，知覚年齢スコアを変化させて生成される変換音声に対して，知覚年齢の付与を行う．3.2 節で述べた個人性を保持する声質制御法（Modified MR-GMM）において，差分知覚年齢スコアを -60, -40, -20, 0, 20, 40, 60 として変換音声を生成する．

3.1 節で述べた従来の MR-GMM に基づく声質制御法（Conventional MR-GMM）と個人性を保持する声質制御法との比較を行うため，変換音声の個人性に関する評価を行う．前実験と同様に評価歌手と被験者を 2 グループに分けて実験を行う．評価は XAB テストにより行い，評価歌手の同一歌手 SVC 歌声を参照音声として被験者に提示した後に，2 手法による変換音声をランダムな順番で提示する．被験者は，どちらの変換音声が参照音声と類似した個人性を持っているかという基準で評価を行う．差分知覚年齢スコアを -60, -30, 30, 60 として変換音声を生成する．従来の MR-GMM に基づく声質制御法に対しては，知覚年齢スコアを同一歌手 SVC 歌声（前実験において差分知覚年齢スコアを 0 とした際）の知覚年齢スコアを基準に $\pm 30, 60$ して，変換音声を生成する．

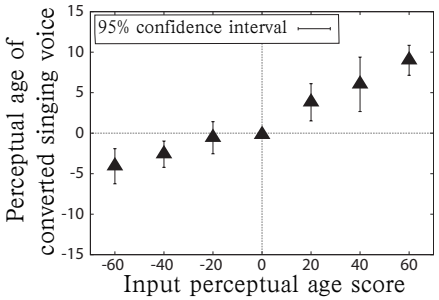

Fig. 1　指定した差分知覚年齢スコアと変換歌声の知覚年齢

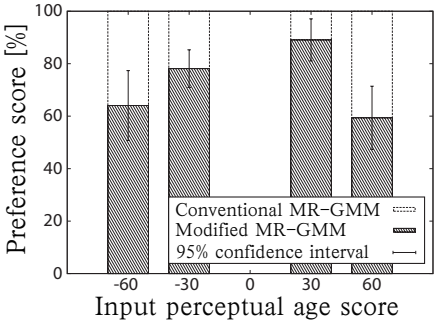

Fig. 2　個人性に関する対比較実験結果

4.2 実験結果

図 1 に知覚年齢に基づく歌声声質制御の精度に関する評価結果を示す．横軸は，指定した差分知覚年齢スコアを表す．縦軸は，被験者が変換音声に対して付与した知覚年齢と，同一歌手 SVC 歌声の知覚年齢との変化量を表す．各点は，評価歌手毎に変化量を計算し，差分知覚年齢スコア別に平均化した値を示す．実験結果より，提案法により，知覚年齢に基づく歌声声質制御が可能であることが分かる．

図 2 に変換音声の個人性に関する Modified MR-GMM と Conventional MR-GMM の比較結果を示す．Modified MR-GMM は Conventional MR-GMM に比べ，歌手の個人性を保持した知覚年齢制御が可能であることが分かる．

5 まとめ

歌声声質変換において，重回帰混合正規分布モデルに基づく声質制御を適用し，知覚年齢に沿った歌声声質制御法を提案した．また，声質制御対象歌手の個人性を保持した知覚年齢操作を可能とするための手法を提案した．実験結果より，提案手法は個人性を保持したまま，目標歌手の知覚年齢を制御可能であることを示した．今後，音声品質に関する評価や，変換音声の高品質化に取り組む予定である．

謝辞　本研究の一部は，JSPS 科研費 22680016 および JST OngaCREST プロジェクトの助成を受け実施したものである．

参考文献

[1] H. Doi *et al.*, Proc. APSIPA ASC, 2012.
[2] K, Ohta *et al.*, Proc. INTERSPEECH, pp.2438–2441, 2010.
[3] T. Toda *et al.*, IEEE Trans. ASLP, Vol. 15, No. 8, pp. 2222–2235, 2007.
[4] 小林和弘 他, 情報処理学会研究報告, Vol.2013–MUS–99 No.44, pp. 1–6, 2013.
[5] 後藤真孝 他, 情報処理学会研究報告, Vol. 2005–MUS–61–2, No. 82, pp. 7–12, 2005.
[6] H. Kawahara *et al.*, Speech Communication, Vol. 27, No. 3–4, pp. 187–207, 1999.