性別依存重回帰混合正規分布モデルに基づく差分スペクトル補正による歌声の知覚年齢制御法*

☆小林 和弘，戸田 智基（奈良先端大・情報），中野 倫靖，後藤 真孝（産総研），
ニュービッグ グラム，サクリアニ サクティ，中村 哲（奈良先端大・情報）

1 はじめに

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できるが，身体的制約を超えた歌唱は困難である．この制約を取り除く一つの方法として，重回帰混合正規分布モデル（MR-GMM: Multiple-Regression Gaussian Mixture Model）[1] に基づく個人性を保持した知覚年齢制御法 [2] が提案されている．しかし，従来の知覚年齢制御法では，知覚年齢の制御範囲が狭いことや変換歌声の自然性の劣化などの問題が生じる．

本稿では，歌声の知覚年齢制御において知覚年齢変換精度と変換歌声の自然性を向上する手法を提案する．まず，従来の知覚年齢制御法に対し，男女間における年齢の変化に伴うスペクトル特徴量の違いを考慮した性別依存 MR-GMM を導入する．さらに，差分スペクトル補正に基づく歌声声質変換（SVC: Singing Voice Conversion）の枠組みを知覚年齢制御法に適用する．実験結果より，提案法は高い自然性を保ちつつ，知覚年齢制御範囲を広げることが可能であることを示す．

2 個人性を保持した知覚年齢制御

個人性を保持した歌声の知覚年齢制御法 [2] では，まず，一人の参照歌手と複数の事前収録目標歌手が同一楽曲を歌唱したパラレルデータを用いて，次式の MR-GMM を学習する．

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

ここで，$\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ と $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^{(s)\top}, \Delta\boldsymbol{y}_t^{(s)\top}\right]^\top$ は，参照歌手と s 番目の事前収録目標歌手の静的・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．α_m は m 番目の分布の混合重みである．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．

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

ここで，$\boldsymbol{b}_m^{(Y)}$ 及び $\overline{\boldsymbol{\mu}}_m^{(Y)}$ は，知覚年齢スコアに対する代表ベクトル及び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知覚年齢スコアを表す．

任意の入力歌手に対して個人性を保持した知覚年齢制御を実現するために，MR-GMM に対し，多対多 SVC の枠組み [3] と平均ベクトルの修正を適用する．その結果，次式で与えられる多対多修正 MR-GMM が得られる．

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

$$\boldsymbol{\mu}_m^{(Y)}(o) = \hat{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{b}_m^{(Y)} \Delta w \quad (4)$$

$$\boldsymbol{\Sigma}_m^{(YXY)} = \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)^{-1}} \boldsymbol{\Sigma}_m^{(XY)} \quad (5)$$

ここで，$\hat{\boldsymbol{\mu}}_m^{(Y)}$ は，参照歌手と入力歌手のパラレルデータを用いて最尤推定される入力歌手の平均ベクトルである．Δw は，入力歌手の知覚年齢を基準とした際の，出力歌手の差分知覚年齢スコアを表す．

変換処理では，所望の差分知覚年齢スコア Δw を与えた多対多修正 MR-GMM に基づき，最尤系列変換法 [4] により，入力歌声を所望の知覚年齢へと変換する．

3 性別依存 MR-GMM に基づく差分スペクトル補正による知覚年齢制御

3.1 性別依存 MR-GMM

従来の知覚年齢制御法では，参照歌手と全ての事前収録目標歌手のデータを用いて一つの MR-GMM を学習する．しかし，男女間において，年齢に伴うスペクトル特徴量の変動には違いがあることが報告されており [5]，従来法では十分なモデル化精度が得られてない可能性がある．そこで，事前収録目標歌手を各性別の 2 つに分割し，性別依存 MR-GMM をそれぞれ学習する．

3.2 差分スペクトル補正に基づく知覚年齢制御

変換歌声の自然性を向上するために，差分スペクトル補正に基づく SVC[6] の枠組みを個人性を保持した知覚年齢制御法に適用する．多対多修正 MR-GMM から得られる差分スペクトル補正用 MR-GMM は以下の式で表される．

$$P\left(\boldsymbol{Y}_t^{(i)}, \boldsymbol{D}_t | \boldsymbol{\lambda}^{(MR)}, \hat{\boldsymbol{\mu}}_m^{(Y)}, \Delta w\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{Y}_t^{(i)} \\ \boldsymbol{D}_t\end{bmatrix}; \begin{bmatrix}\hat{\boldsymbol{\mu}}_m^{(Y)} \\ \boldsymbol{b}_m^{(Y)} \Delta w\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(YY)} \boldsymbol{\Sigma}_m^{(DYD)} \\ \boldsymbol{\Sigma}_m^{(DYD)} \boldsymbol{\Sigma}_m^{(DD)}\end{bmatrix}\right) \quad (6)$$

$$\boldsymbol{\Sigma}_m^{(DYD)} = \boldsymbol{\Sigma}_m^{(YXY)} - \boldsymbol{\Sigma}_m^{(YY)} \quad (7)$$

$$\boldsymbol{\Sigma}_m^{(DD)} = 2(\boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(YXY)}) \quad (8)$$

ここで，$\boldsymbol{D}_t = \left[\boldsymbol{d}_t^\top, \Delta\boldsymbol{d}_t^\top\right]^\top$ は，静的・動的差分スペクトル特徴量である．

---

*Improvement of Perceived Age Control in Singing Voice Conversion using Gender-Dependent Modeling and Direct Waveform Modification, by KOBAYASHI, Kazuhiro, TODA, Tomoki（NAIST）, NAKANO, Tomoyasu, GOTO, Masataka（AIST）, NEUBIG, Graham, SAKTI, Sakriani, NAKAMURA, Satoshi（NAIST）

変換処理では，2 節と同様の方法で，差分スペクトル補正用 MR-GMM に基づき，差分スペクトル特徴量系列が推定される．入力自然歌声に対して，差分スペクトル特徴量系列を畳み込むことで，所望の知覚年齢を持つ変換歌声が得られる．なお，本稿では，差分スペクトル特徴量の系列内変動 [4] については考慮しない．

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 [7] を用いる．評価楽曲は No.39 とする．性別非依存 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．また，性別依存 MR-GMM の学習では，男女別に実年齢が 20 代の参照歌手 1 名と事前収録目標歌手 27 名をそれぞれ用いる．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）を評価歌手として用いる．被験者は 20 代男性 8 名である．評価歌手と被験者は，各性別・年代の人数が均一となる様に 2 グループに分け評価を行う．

スペクトル包絡パラメータとして，STRAIGHT 分析 [8] 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．

実験では，差分知覚年齢スコアを-60, -30, 0, 30, 60 に設定し，以下のモデルを用いて得られる変換歌声を評価する．

- SVC (I): 性別非依存 MR-GMM
- SVC (D): 性別依存 MR-GMM
- DIFFSVC (D): 性別依存差分 MR-GMM

知覚年齢変換精度を評価するために，知覚年齢スコアを基づき生成される変換歌声に対して，知覚年齢の評価を行う．SVC（I），SVC（D），DIFFSVC (D) の変換歌声を生成し，被験者はランダムな順番再生される変換歌声に対し知覚年齢を評価する．

変換歌声の自然性を，5 段階評定オピニオン（MOS: Mean Opinion Score）に基づき評価する．被験者に対して，自然歌声，SVC（I），SVC（D），DIFFSVC (D) 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．

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，差分知覚年齢スコアを表す．縦軸は，差分知覚年齢スコアを 0 に設定した時の知覚年齢と各差分知覚年齢スコアにおける変換歌声の知覚年齢の変化量を表す．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，性別依存 MR-GMM は性別非依存 MR-GMM に比べ．より広い範囲で知覚年齢を制御出来る事がわかる．

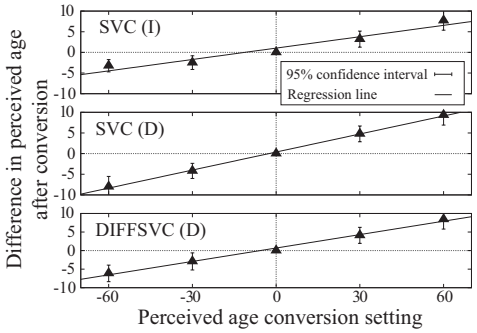

Fig. 1　差分知覚年齢スコアと変換歌声の知覚年齢

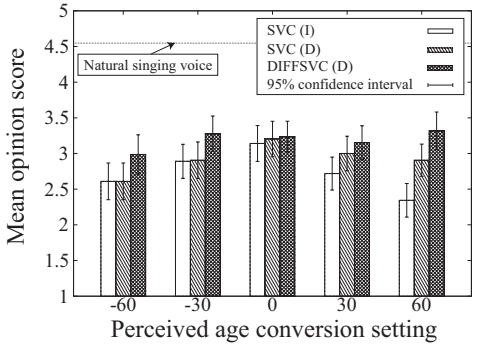

Fig. 2　自然性に関する MOS 評価

図 2 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．性別依存 MR-GMM により，自然性の向上が確認される．さらに，差分スペクトル補正により更なる自然性の向上が確認される．

5 まとめ

重回帰混合正規分布モデルに基づく歌声の知覚年齢制御法において，性別依存モデルと差分スペクトル補正を適用する手法を提案した．実験結果より，提案法は従来法に比べ，高い自然性を保ちつつより広い範囲での知覚年齢制御を実現できることを示した．今後，適応データ数を減らした際の変換精度の評価や，年齢変動に影響を与える差分スペクトル特徴量の分析を行う．

謝辞　本研究の一部は，JSPS 科研費 26280060 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] K, Ohta *et al.*, Proc. INTERSPEECH, pp.2438–2441, 2010.
[2] K, Kobayashi *et al.*, IEICE Trans. Inf. Syst., Vol. E97–D, No. 6, pp. 1419–1428, 2014.
[3] H. Doi *et al.*, Proc. APSIPA ASC, 2012.
[4] T. Toda *et al.*, IEEE Trans. ASLP, Vol. 15, No. 8, pp. 2222–2235, 2007.
[5] S. E. Linville *et al.*, Journal of Voice, Vol. 15, Npp. 3, pp. 323–330, 2001.
[6] K, Kobayashi *et al.*, Proc. INTERSPEECH, 2014.
[7] 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.
[8] H. Kawahara *et al.*, Speech Communication, Vol. 27, No. 3–4, pp. 187–207, 1999.